mathstat2020-ZH

目录

第一章引言

统计学在概率论起作用的场景下,研究利用随机模型对可观测现象进行建模，并分析其产生数据的方式：常用的方法有估计模型参数、构造置信区间以及假设检验.

在这份讲义中，我们将探讨各种估计和检验方法，考察它们的理论性质，并探讨不同的最优性概念.

一些符号和模型假设

数据由测量值（观测值）$x_1, \ldots, x_n$ 组成，这些值被视为随机变量 $X_1, \ldots, X_n$ 的实现.在大部分内容中，X_i 是(一维)实值变量，即 $X_i \in \mathbb{R}$（$i = 1, \ldots, n$），尽管我们也会考虑向量值观测的一些扩展.

1.1 光速例子

费佐和傅科在1849年和1850年分别开发了估计光速的方法，这些方法后来被纽康和迈克尔逊改进.基本思想是让光从一个快速旋转的镜子反射到固定镜子上，再返回到旋转镜子.通过结合旋转镜子的速度、光程距离和返回时光的位移，可以估计光速.

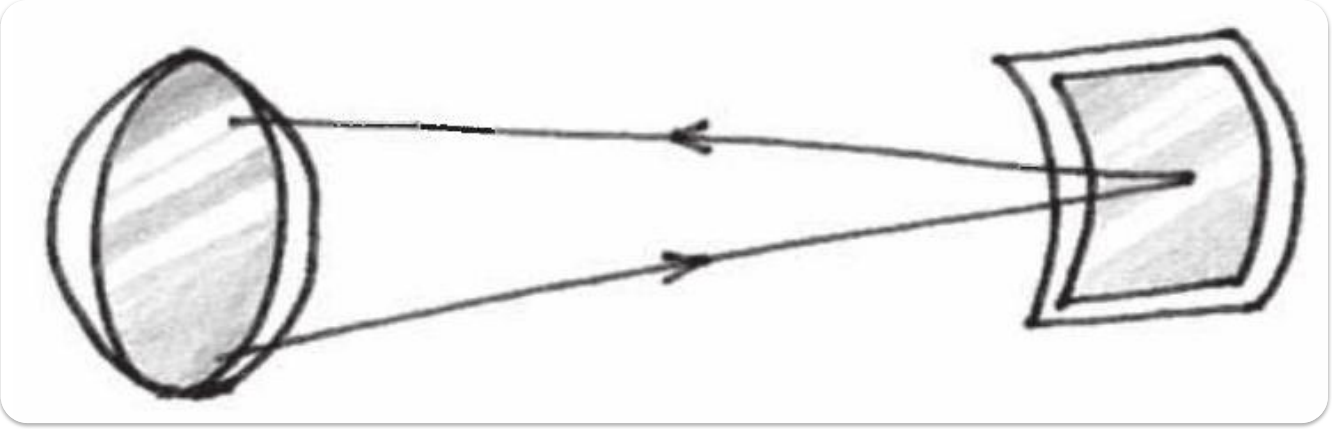
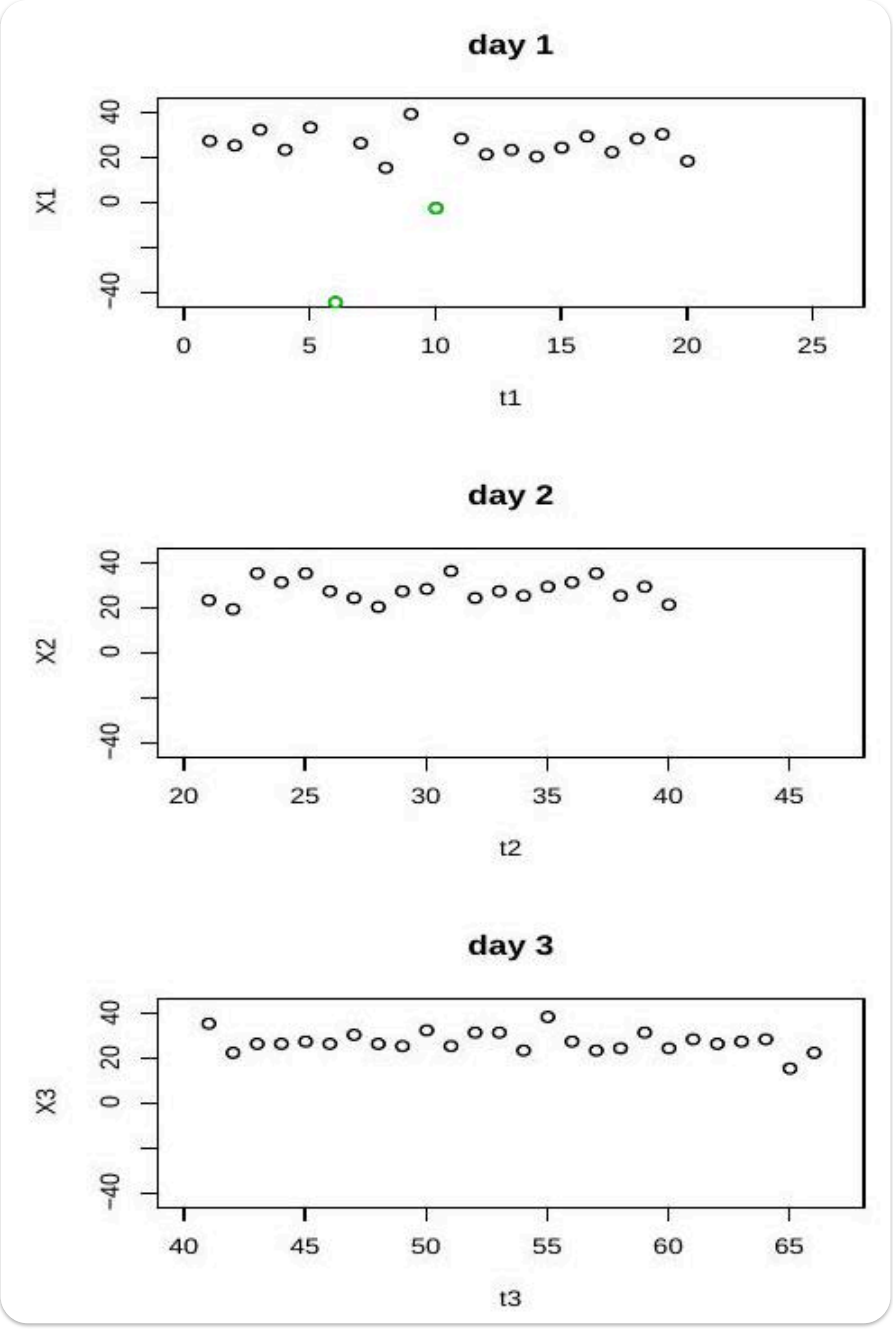


图1
数据来源于纽康对光从实验室到华盛顿纪念碑镜子再返回实验室的光程时间的测量：

- 距离：7.44373公里
- 连续三天的66次测量
- 第一次测量：0.000024828秒 = 24828 纳秒
- 数据集中记录的是相对于24800纳秒的偏差值

三天的测量结果如下表所示：



可以通过均值、中位数或Huber估计（如下所述）的方法分别估算光速.对于三天的单独测量以及三天的综合数据，这些估计的结果如下：

方法	第一天	第二天	第三天	总体数据
均值	21.75	28.55	27.85	26.21
中位数	25.5	28	27	27
Huber估计	25.65	28.40	27.71	27.28

表 1

“哪个估计是‘最佳’的”是我们讲义探讨的主题之一.

1.2 统计模型和分布函数

观测值的集合记为 $\mathbf{X} = \{X_1, \ldots, X_n\}$.$\mathbf{X}$ 的分布记为 $\mathbb{P}$，通常是未知的.我们定义统计模型为:

统计模型

统计模型是给定样本空间上一组概率分布的集合.

统计模型的解释是：观察值 X 的所有可能的概率分布的集合. 通常，观察值由“子观察值”组成，$X = (X_1, \ldots, X_n)$ 是一个随机向量. 当变量 $X_1, \ldots, X_n$ 对应于相同实验的独立重复时，我们称之为样本. 这时，变量 $X_1, \ldots, X_n$ 是独立同分布的，它们的联合分布完全由相同的边际分布决定. 在这种情况下，$X = (X_1, \ldots, X_n)$ 的统计模型可以通过子观察值 $X_1, \ldots, X_n$ 的一组（边际）概率密度函数来描述 (因为每一个都是独立同分布, 一定要用联合分布可能显得麻烦了).

这一分布，即X的分布，用P表示.对于$X \in \mathbb{R}$，X的分布函数表示为:

分布函数

$$F(\cdot) = P(X \leq \cdot)$$

请注意，分布函数F决定了分布P，反之亦然.

进一步的模型假设则涉及对 P 的建模.我们将模型写为 $P \in \mathcal{P}$，其中 $\mathcal{P}$ 是一组概率测度，称为模型类.通常，$\mathcal{P}$ 中的分布以某个参数（例如 θ）在某参数空间（例如 Θ）内的取值作为指标.

这样，我们可以如下表示 $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$，并且我们说真实分布 $P = P_\theta$ 对于某个 $\theta \in \Theta$.此时，θ 通常称为“真参数”.

参数空间 Θ 可能是高维的，甚至是无限维的.

通常，人们只对参数的某一特定方面感兴趣.我们将关注的参数的某一方面写为 $\gamma := g(\theta)$，其中 $g : \Theta \to \Gamma$ 是一个定义在某空间 Γ 上的函数.

1.3 例子：位置模型

以下例子非常重要, 会多次用来说明后续的概念.

令X为实值变量，位置模型定义为

位置模型

$$\mathcal{P} := \{P_\theta(X \leq \cdot) := F_0(\cdot - \mu), \theta := (\mu, F_0), \mu \in \mathbb{R}, F_0 \in \mathcal{F}_0\}$$

其中 $\mathcal{F}_0$ 是一组给定的分布函数集合.

如果期望存在，我们将 $\mathcal{F}_0$ 中的分布中心化为零均值.

此时，P_{μ,F_0} 的均值为 μ.我们称 μ 为位置参数.通常，仅 μ 是感兴趣的参数，而 F_0 是所谓的无关参数（nuisance parameter）. 在这种情况下，按照我们之前提到的表达方法, $g(\mu, F_0) = \mu$.

接下来让我们举几个 $\mathcal{F}_0$ 的例子.

对称分布类

$\mathcal{F}_0$类的一个例子是对称分布类：

$$\mathcal{F}_0 := \{F_0(x) = 1 - F_0(-x), \forall x\}. \tag{1.2}$$

这是一个无限维集合，因为它并非由有限维参数表示.因此，我们称F_0为无限维参数.

正态分布有限维类

一个有限维模型的例子是：

$$\mathcal{F}_0 := \{\Phi(\cdot/\sigma) : \sigma > 0\} \tag{1.3}$$

其中Φ是标准正态分布函数.

因此，位置模型可表示为

$$X_i = \mu + \epsilon_i, i = 1, \ldots, n$$

其中 $\epsilon_1, \ldots, \epsilon_n$ 独立同分布，并且在模型(1.2)下对称分布但其他关于分布的信息未知，而在模型(1.3)下服从方差未知的 $\mathcal{N}\left(0, \sigma^2\right)$ 分布.

1.4 一些统计模型的其他例子

例子 1.4.1 泊松分布

保险公司

考虑一家小型保险公司.令X表示某天的索赔数量.此时可能的模型是泊松模型，假设X服从参数为$\theta > 0$的泊松分布：

$$P_\theta(X = x) = \frac{\theta^x}{x!}\mathrm{e}^{-\theta}, x = 0, 1, 2, \ldots$$

一个感兴趣的参数的某一方面可能是某天至少有4个索赔的概率：

$$\begin{aligned}
\gamma &= P_\theta(X \geq 4) \\
&= 1 - P_\theta(X \leq 3) \\
&= 1 - \left(1 + \theta + \frac{\theta^2}{2} + \frac{\theta^3}{3!}\right)\mathrm{e}^{-\theta} \\
&:= g(\theta).
\end{aligned}$$

假设我们观测了 $n = 200$ 天的索赔数量 $X_1, \ldots, X_n$.
θ 的一个可能的也是很直观的估计量是样本均值 $\bar{X} := \sum_{i=1}^{n} X_i/n$.
对于 γ，我们可以使用“代入”原则，即将 $\bar{X}$ 代入 g 函数中的 θ.
这样，$g(\bar{X})$ 作为 $g(\theta)$ 的估计量记为 $\hat{\gamma} := g(\bar{X})$.

以下是一个数据例子：

x_i	**天数**
0	100
1	60
2	32
3	8
≥ 4	0

观测到的平均值为 $\bar{x} = 0.74$，那么估计的 $P_\theta(X \geq 4)$ 为 0.00697.

例子 1.4.2 帕累托分布

收入分布

假设 X 的密度函数（相对于 Lebesgue 测度 ν）为：

$$p_\theta(x) = \theta(1+x)^{-(1+\theta)},\ x > 0$$

其中 $\theta > 0$ 是未知参数.

这是一个帕累托分布的密度，常用于建模收入分布.参数 θ 有时称为帕累托指数.

这个时候分布的模型类可以表示为：

$$\mathcal{P} = \left\{ P_\theta : \frac{dP_\theta}{d\nu} = p_\theta \right\}$$

一个感兴趣的参数是基尼系数（Gini index），它描述了收入不平等.
在 $\theta \geq 1$ 的情况下定义为 $\gamma(\theta) = 1/(2\theta - 1)$.
当 $\theta = 1$ 时，基尼系数为 $G(1) = 1$，表示完全收入不平等.

例子 1.4.3 分类问题

糖尿病概率

设 $X = (Y, Z)$，其中 Z 表示体重指数（BMI），$Z \in \mathbb{R}$，$Y \in \{0, 1\}$ 是一个指标参数,表示是否患有糖尿病.

假设模型为：

$$P_\theta(Y = 1 \mid Z = z) = \theta(z),\ z \in \mathbb{R}$$

其中

$$\theta(\cdot) \in \Theta := \{\text{所有递增函数 } \theta : \mathbb{R} \to [0,1]\}$$

可见这里参数空间是 ∞ 维的.

一个感兴趣的参数是:

$$\gamma := \theta^{-1}\left(\frac{1}{2}\right),$$

即满足以下条件的最小值 γ:

$$P_\theta(Y = 1 \mid Z = z) \geq \frac{1}{2},\ z \geq \gamma$$

例子 1.4.4 社交网络

个体联系

考虑 p 个个体，他们之间可能存在或不存在联系.如果存在联系，我们称他们为朋友，或者说他们之间有连接.

令 X 为 $p \times p$ 的矩阵 $X = (X_{j,k})$，用于编码连接关系:

对于 $j \neq k$,

$$X_{j,k} := \begin{cases} 1 & \text{如果 } j \text{ 和 } k \text{ 之间有连接} \\ 0 & \text{否则} \end{cases}$$

"随机块模型"（stochastic block model）假设 $\{X_{j,k} : j < k\}$ 是独立的，且

$$P_\theta\left(X_{j,k} = 1\right) = \begin{cases} \beta_m & \text{如果 } j \text{ 和 } k \text{ 属于同一社区 } m \\ \delta & \text{如果 } j \text{ 和 } k \text{ 属于不同社区} \end{cases}$$

如果社区的数量已知（设为 M），但没有进一步信息，则参数空间为:

$$\Theta = \left\{(\beta_1, \ldots, \beta_M, \delta) \in [0,1]^{M+1}, \mathcal{M}\right\}$$

其中 $\mathcal{M}$ 是所有人所属社区的所有可能性的集合.

共有 M^p 种可能的人员社区配置，因此 $|\mathcal{M}| = M^p$.
当 p 较大时，这是一个极为复杂的参数空间.

备注：通常我们只能观测到 X (人际关系)的一个实现, 不太可能有多个数据，即样本数量 $n = 1$.

例子 1.4.5 因果模型

发现地区参数的因果关系

假设 $X=(Z_1,\ldots,Z_p)\in\mathbb{R}^p$ 是一个 p 维随机变量，例如 Z_1 表示降雨量，Z_2 表示茶叶消费量，Z_3 表示高个子人数，Z_4 表示山峰高度，等等，这些数据来自某个国家的某个特定地区.因果模型旨在找出哪些变量是因，哪些是果.

结构关系模型为：

$$Z_{\pi(j)}=f_j\left(Z_{\pi(1)},\ldots,Z_{\pi(j-1)},\epsilon_j\right),\ j=2,\ldots,p$$

其中 π 是 $1,\ldots,p$ 的一个排列，ϵ_j 是不可观测的噪声，f_j 是部分未知的函数.

如果假设（为简单起见）噪声分布已知，则参数空间为：

$$\Theta:=\{\text{所有排列 }\pi\text{ 和结构关系 }(f_2,\ldots,f_p)\}$$

一个感兴趣的参数是因果图.
一个具体形式是将结构关系建模为线性关系：

$$f_j\left(z_1,\ldots,z_{j-1},\varepsilon_j\right)=\beta_{j,1}z_1+\cdots+\beta_{j,j-1}z_{j-1}+\varepsilon_j,\ j=2,\ldots,p$$

其中 $\{\beta_{j,k}\}$ 是未知系数.

第二章参数估计

2.1 什么是估计量?

回忆一下，数据由观测值 $\mathbf{X}=(X_1,\ldots,X_n)$ 组成，其分布部分未知.
参数是未知分布的某个方面.
我们通常假设 $X_1,\ldots,X_n$ 是随机变量 X 的独立同分布 (i.i.d.) 副本，其中 X 的分布为 $P_\theta\in\{P_\vartheta:\vartheta\in\Theta\}$.
构造估计量的目的是估计某个未知参数 γ.其形式定义如下：

估计量

估计量 $T(\mathbf{X})$ 是某个给定（可测）函数 $T(\cdot)$ 在观测值 $\mathbf{X}$ 上的取值.
函数 $T(\cdot)$ 不允许依赖未知参数.

估计量也被称为统计量或决策量.
T 不允许依赖未知参数的原因是，估计量在实际应用中应该只依赖数据进行计算.
我们经常用相同的符号 T 来表示估计量 $T(\mathbf{X})$（即，我们写作 $T=T(\mathbf{X})$，省略了参数 $\mathbf{X}$）和函数 $T=T(\cdot)$.

2.2 经验分布函数

令 $X_1,\ldots,X_n$ 为实值观测值.非参数估计量的一个例子是经验分布函数：

经验分布函数

$$\hat{F}_n(\cdot) := \frac{1}{n}\#\{X_i \le \cdot, 1 \le i \le n\}$$

这是理论分布函数

$$F(\cdot) := P(X \le \cdot)$$

的一个估计量.

大多数估计量是根据所谓的"插入"原则（plug-in principle, 或 Einsetzprinzip）构造的.即，感兴趣的参数 γ 被写作 $\gamma = Q(F)$，其中 Q 是某个给定的映射.然后将经验分布函数 $\hat{F}_n$ "插入", 作为真实分布的一种"替代"，得到可以用来估计 $\gamma = Q(F)$ 的估计量 $T := Q\left(\hat{F}_n\right)$.
（然而我们注意到可能会出现问题，例如 $Q\left(\hat{F}_n\right)$ 可能无法定义 ……）.

2.3 关于位置模型的一些估计量

在第 1.3 节的位置模型中，可以考虑以下 μ 的估计量 $\hat{\mu}$（以及其他估计量）：

- **均值**

$$\hat{\mu}_1 := \frac{1}{n}\sum_{i=1}^{n} X_i$$

注意，$\hat{\mu}_1$ 最小化平方损失 [1]

$$\sum_{i=1}^{n}(X_i - \mu)^2$$

即

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n}(X_i - \mu)^2. \tag{2.1}$$

可以证明，当模型 (1.3) 成立时（即 $X_1, \ldots, X_n$ 是独立同分布的 $\mathcal{N}\left(\mu, \sigma^2\right)$ 随机变量），$\hat{\mu}_1$ 是一个“好”的估计量.当模型 (1.3) 不成立，特别是存在离群值（过大或“错误”的观测值）时，我们就需要使用更稳健的估计量.

- **（样本）中位数**

中位数

$$\hat{\mu}_2 := \begin{cases} X_{((n+1)/2)} & \text{当 } n \text{ 为奇数时} \\ \left\{X_{(n/2)} + X_{(n/2+1)}\right\}/2 & \text{当 } n \text{ 为偶数时} \end{cases}$$

其中 $X_{(1)} \le \cdots \le X(n)$ 为顺序统计量.注意，$\hat{\mu}_2$ 最小化**绝对损失**

$$\sum_{i=1}^{n}|X_i-\mu|$$

- Huber 估计量1

$$\hat{\mu}_3:=\arg\min_{\mu\in\mathbb{R}}\sum_{i=1}^{n}\rho\left(X_i-\mu\right)$$

其中

$$\rho(x)=\left\{\begin{array}{ll}x^2 & \text{如果 } |x|\leq k\\ k(2|x|-k) & \text{如果 } |x|>k\end{array}\right\}$$

$k>0$ 是一个给定的阈值.

- α-截断均值

对于某些 $0<\alpha<1$，定义为

$$\hat{\mu}_4:=\frac{1}{n-2[n\alpha]}\sum_{i=[n\alpha]+1}^{n-[n\alpha]}X_{(i)}$$

上述估计量 $\hat{\mu}_1,\ldots,\hat{\mu}_4$ 是 μ 的"插入"估计量.

结合之前说的, 给分布一个映射以聚焦于我们关心的参数,我们定义以下映射:

$Q_1(F):=\int xdF(x)$（F 的均值, 期望或重心），

$Q_2(F):=F^{-1}(1/2)$（F 的中位数），

$Q_3(F):=\arg\min_{\mu}\int\rho(\cdot-\mu)dF$,

以及 $Q_4(F):=\frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)}xdF(x)$.

则 $\hat{\mu}_k$ 对应于 $Q_k\left(\hat{F}_n\right),k=1,\ldots,4.$ 是对于几种不同的关心的参数的估计 (量).

这几种估计量对于分布是否对称有不同的意义:
如果模型 (1.2) 成立（即 F 关于 μ 对称），则 $\hat{\mu}_1,\ldots,\hat{\mu}_4$ 都是 μ 的估计量. 直观来说可以想象一个对称分布的期望, 中位数, Huber 数, α-截断的图像所在, 容易发现这是重叠的.
如果对称模型假设不成立，则每个 $Q_k\left(\hat{F}_n\right)$ 仍是 $Q_k(F)$ 的一个估计量（假设后者存在），但 $Q_k(F)$ 可能就是 F 的不同方面了.

2.4 如何构造估计量

2.4.1 "插入"估计量

对于实值观测，可以定义分布函数

$$F(\cdot)=P(X\leq\cdot)$$

之前就讲到了分布函数的一个估计量是经验分布函数, 现在我们换一种方式表达经验分布:

经验分布 (示性函数表达)

$$\hat{F}_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq \cdot\}$$

注意，当仅知道 $\hat{F}_n$ 时，可以据此构建顺序统计量 $X_{(1)} \leq \ldots \leq X_{(n)}$，因为我们的经验分布函数会在每一个数据点的地方跳跃一下,但同样显然的是我们不能还原原始数据 $X_1, \ldots, X_n$.

不过到底能不能还原原始数据并不是一件重要的事情.事实上，数据的排列顺序不携带关于分布 P 的信息.换句话说，一个“合理的” [2] 估计量 $T = T(X_1, \ldots, X_n)$ 仅通过样本的顺序统计量 $\left(X_{(1)}, \ldots, X_{(n)}\right)$ 依赖于样本（即，重新排列数据不应影响 T 的值）.

由于可以从经验分布 $\hat{F}_n$ 确定这些顺序统计量，我们可以得出结论，任何“合理的”估计量 T 都可以表示为 $\hat{F}_n$ 的一个函数:

$T = Q\left(\hat{F}_n\right)$,其中 Q 是某个映射.

类似地，分布函数 $F_\theta := P_\theta(X \leq \cdot)$ 完全刻画了分布 P_θ.
因此，参数可以表示为 F_θ 的一个函数:

$\gamma = g(\theta) = Q(F_\theta)$.

如果映射 Q 在所有 F_θ 和 $\hat{F}_n$ 上都定义，则称 $Q\left(\hat{F}_n\right)$ 是 $Q(F_\theta)$ 的一个"插入"估计量.

这一思想并不限于一维情况.对于任意观测空间 $\mathcal{X}$，定义经验测度

$$\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

其中 δ_x 是"点质量"在 x 的一个 Dirac 测度.经验测度在每个观测值上赋予"质量" $1/n$.这实际上是 $\mathcal{X} = \mathbb{R}$ 的推广，因为经验分布函数 $\hat{F}_n$ 在每个观测值处跃升，跃升高度等于该值的观测次数（如果所有 X_i 不同，则每一个的跃升高度都为 $1/n$）.与实值情况类似，如果映射 Q 在所有 P_θ 和 $\hat{P}_n$ 上都定义，则称 $Q\left(\hat{P}_n\right)$ 是 $Q(P_\theta)$ 的一个"插入"估计量.

需要强调的是，通常 $\gamma = g(\theta)$ 作为 P_θ 的函数 Q 的表示形式并不唯一，即可以有多种选择 Q. 每种选择通常会导致不同的估计量.

此外，假设 Q 在 $\hat{P}_n$ 上有定义这一前提也常常不一定满足.有时可以将映射 Q 修改为 Q_n，使其在某种意义上对大样本 n 逼近 Q.此时，修正后的"插入"估计量形式为 $Q_n\left(\hat{P}_n\right)$.

2.4.2 矩估计法

设 $X \in \mathbb{R}$，假设感兴趣的参数是 θ 本身，并且 $\Theta \subset \mathbb{R}^p$.

设 X 的前 p 阶矩（假设存在）为 $\mu_1(\theta),\ldots,\mu_p(\theta)$，即
$\mu_j(\theta)=\mathbb{E}_\theta X^j=\int x^j dF_\theta(x),\quad j=1,\ldots,p.$ 说到矩我们常常会说原点矩和中心矩两种，这里的矩特指的是原点矩，之后不再赘述.

我们还假设映射: $m:\Theta\to\mathbb{R}^p$.由下式定义: $m(\theta)=[\mu_1(\theta),\ldots,\mu_p(\theta)]$
并且该映射在所有 $[\mu_1,\ldots,\mu_p]\in\mathcal{M}$（假设集合）上存在逆映射
$m^{-1}(\mu_1,\ldots,\mu_p).$

我们用样本矩来估计 μ_j，即
$\hat{\mu}_j:=\frac{1}{n}\sum_{i=1}^n X_i^j=\int x^j d\hat{F}_n(x),\quad j=1,\ldots,p.$

当 $\left[\hat{\mu}_1,\ldots,\hat{\mu}_p\right]\in\mathcal{M}$ 时，可以将其代入，得到估计量
$\hat{\theta}:=m^{-1}\left(\hat{\mu}_1,\ldots,\hat{\mu}_p\right).$

例子 2.4.1 矩估计法在负二项分布中的应用

设 X 服从具有已知参数 k 和未知成功概率 $\theta\in(0,1)$ 的负二项分布:

$$P_\theta(X=x)=\binom{k+x-1}{x}\theta^k(1-\theta)^x,\quad x\in\{0,1,\ldots\}$$

这是在独立试验中达到第 k 次成功之前失败的次数的分布，其中每次试验成功的概率为 θ.其期望为

$$\mathbb{E}_\theta(X)=k\frac{(1-\theta)}{\theta}:=m(\theta)$$

因此$m^{-1}(\mu)=\frac{k}{\mu+k}$，

矩估计量为

$$\hat{\theta}=\frac{k}{\bar{X}+k}=\frac{nk}{\sum_{i=1}^n X_i+nk}=\frac{\text{成功次数}}{\text{试验次数}}$$

例子 2.4.2 矩估计法在帕累托分布中的应用

假设 X 的密度函数为$p_\theta(x)=\theta(1+x)^{-(1+\theta)},\quad x>0$，

其对应的勒贝格测度，且 $\theta\in\Theta\subset(0,\infty)$（见例子 1.4.3）.

当 $\theta>1$ 时，$\mathbb{E}_\theta X=\frac{1}{\theta-1}:=m(\theta)$，

其逆映射为$m^{-1}(\mu)=\frac{1+\mu}{\mu}$.

因此，矩估计量为$\hat{\theta}=\frac{1+\bar{X}}{\bar{X}}$.

然而，当 $\theta<1$ 时，期望 $\mathbb{E}_\theta X$ 不存在，因此当 Θ 包含 $\theta<1$ 的取值时，矩估计法可能不是一个好的选择.我们将从后面的章节看到最大似然估计不会有这个问题.

2.4.3 似然方法

假设 $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ 被一个 σ -有限测度 ν 主导.
我们将密度函数表示为

$$p_\theta := \frac{dP_\theta}{d\nu}, \quad \theta \in \Theta$$

似然函数

似然函数（基于数据 $\mathbf{X} = (X_1, \ldots, X_n)$）是一个函数 $L_{\mathbf{X}} : \Theta \to \mathbb{R}$，其定义为

$$L_{\mathbf{X}}(\vartheta) := \prod_{i=1}^{n} p_\vartheta(X_i), \quad \vartheta \in \Theta$$

最大似然估计量（MLE）

$$\hat{\theta} := \arg\max_{\vartheta \in \Theta} L_{\mathbf{X}}(\vartheta)$$

注释

注释 我们用符号 ϑ 表示似然函数中的变量，而用稍微不同的符号 θ 表示我们希望估计的参数.然而，通常的习惯是对两者使用相同的符号（如在 1.2 和 2.3 节脚注中已经提到）.但在理论发展中，使用不同的符号是必要的.

注释 或者，我们可以将 MLE 写为对对数似然的最大化：

$$\hat{\theta} = \arg\max_{\vartheta \in \Theta} \log L_{\mathbf{X}}(\vartheta) = \arg\max_{\vartheta \in \Theta} \sum_{i=1}^{n} \log p_\vartheta(X_i)$$

对数似然通常在数学上更易处理.例如，如果密度函数是可微的，可以通过设置导数为零来得到最大值，并且求和比求积更易于微分.

注释 似然函数可能有局部最大值, 所以仅仅求一阶导看到一个解就认为其是 MLE 是不恰当的.此外，MLE 并不总是唯一的，甚至可能不存在（例如，当似然函数无界时）.

现在我们将证明最大似然是一种也是一种替代估计法.
首先，如上所述，MLE 最大化对数似然.
当然，我们可以通过 $1/n$ 对对数似然进行归一化, 这完全不影响最后的解：

$$\hat{\theta} = \arg\max_{\vartheta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\vartheta(X_i) \tag{2.2}$$

将 (2.2) 中的平均值 $\sum_{i=1}^{n} \log p_\vartheta(X_i)/n$ 替换为其理论对应值 $\mathbb{E}_\theta \log p_\vartheta(X)$，得到

$$\arg\max_{\vartheta \in \Theta} \mathbb{E}_\theta \log p_\vartheta(X)$$

这确实等于我们试图估计的参数 θ:

> **引理 2.4.1**
>
> 我们有
>
> $$\theta = \arg\max_{\vartheta\in\Theta} \mathbb{E}_\theta \log p_\vartheta(X).$$

证明. 由不等式 $\log x \le x-1, x>0$，对所有 $\vartheta \in \Theta$ 有

$$\mathbb{E}_\theta \log \frac{p_\vartheta(X)}{p_\theta(X)} \le \mathbb{E}_\theta \left(\frac{p_\vartheta(X)}{p_\theta(X)} - 1 \right) = 0$$

2.4.4 最大似然估计法 (MLE) 的例子

例子 2.4.3 Pareto 分布的 MLE

假设 X 的密度为

$$p_\theta(x) = \theta(1+x)^{-(1+\theta)}, \quad x>0$$

相对于 Lebesgue 测度，其中 $\theta \in \Theta = (0,\infty)$.则有

$$\log p_\vartheta(x) = \log\vartheta - (1+\vartheta)\log(1+x)$$
$$\frac{d}{d\vartheta}\log p_\vartheta(x) = \frac{1}{\vartheta} - \log(1+x)$$

将基于 n 个观测值的对数似然函数的导数设为零并求解:

$$\frac{n}{\hat\theta} - \sum_{i=1}^n \log(1+X_i) = 0$$
$$\Rightarrow \hat\theta = \frac{1}{\{\sum_{i=1}^n \log(1+X_i)\}/n}$$

（可以验证这是全局最大值.）

例子 2.4.4 某些位置/尺度模型的 MLE

令 $X \in \mathbb{R}$，且 $\theta = (\mu, \sigma^2)$，其中 $\mu \in \mathbb{R}$ 为位置参数，$\sigma > 0$ 为尺度参数.

假设 X 的分布函数为

$F_\theta(\cdot) = F_0\left(\frac{\cdot-\mu}{\sigma}\right)$,

其中 F_0 是相对于 Lebesgue 测度的给定分布函数，其密度为 f_0.因此，X 的密度为

$p_\theta(\cdot) = \frac{1}{\sigma} f_0\left(\frac{\cdot-\mu}{\sigma}\right)$.

情况 1 若 $F_0 = \Phi$（标准正态分布函数），则

$$f_0(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right], \quad x \in \mathbb{R}$$

因此

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad x \in \mathbb{R}$$

μ 和 σ^2 的 MLE 为 (注意这里方差的 MLE 的系数是 1/n)

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

情况 2 （标准化）双指数分布或拉普拉斯分布的密度为

$$f_0(x) = \frac{1}{\sqrt{2}} \exp[-\sqrt{2}|x|], \quad x \in \mathbb{R}$$

因此

$$p_\theta(x) = \frac{1}{\sqrt{2\sigma^2}} \exp\left[-\frac{\sqrt{2}|x-\mu|}{\sigma}\right], \quad x \in \mathbb{R}$$

μ 和 σ 的 MLE 分别为

$$\hat{\mu} = \text{样本中位数}, \quad \hat{\sigma} = \frac{\sqrt{2}}{n}\sum_{i=1}^{n}|X_i - \hat{\mu}|$$

例子 2.4.5 一个 MLE 不存在的例子

以下是 Kiefer 和 Wolfowitz (1956) 提出的一个著名例子，其中似然函数无界，因此 MLE 不存在.

该例涉及两个正态分布的混合：每个观测值或服从 $\mathcal{N}(\mu, 1)$，或服从 $\mathcal{N}\left(\mu, \sigma^2\right)$，两者的概率均为 $1/2$.

未知参数为 $\theta = \left(\mu, \sigma^2\right)$，$X$ 的密度为

$$p_\theta(x) = \frac{1}{2}\phi(x-\mu) + \frac{1}{2\sigma}\phi((x-\mu)/\sigma), \quad x \in \mathbb{R}$$

相对于 Lebesgue 测度.

于是

$$L_{\mathbf{X}}\left(\tilde{\mu}, \tilde{\sigma}^2\right) = \prod_{i=1}^{n}\left(\frac{1}{2}\phi\left(X_i - \tilde{\mu}\right) + \frac{1}{2\tilde{\sigma}}\phi\left(\left(X_i - \tilde{\mu}\right)/\tilde{\sigma}\right)\right)$$

令 $\tilde{\mu} = X_1$，则有

$$L_{\mathbf{X}}\left(X_1, \tilde{\sigma}^2\right) = \frac{1}{\sqrt{2\pi}}\left(\frac{1}{2} + \frac{1}{2\tilde{\sigma}}\right)\prod_{i=2}^{n}\left(\frac{1}{2}\phi\left(X_i - X_1\right) + \frac{1}{2\tilde{\sigma}}\phi\left(\left(X_i - X_1\right)/\tilde{\sigma}\right)\right)$$

由于对于所有 $z \neq 0$ 有

$$\lim_{\tilde{\sigma} \downarrow 0} \frac{1}{\tilde{\sigma}} \phi(z/\tilde{\sigma}) = 0$$

因此

$$\lim_{\tilde{\sigma} \downarrow 0} \prod_{i=2}^{n} \left(\frac{1}{2} \phi\left(X_i - X_1\right) + \frac{1}{2\tilde{\sigma}} \phi\left(\left(X_i - X_1\right)/\tilde{\sigma}\right) \right) = \prod_{i=2}^{n} \frac{1}{2} \phi\left(X_i - X_1\right) > 0$$

于是有$\lim_{\tilde{\sigma} \downarrow 0} L_{\mathbf{X}}\left(X_1, \tilde{\sigma}^2\right) = \infty$.

2.5 基于似然的渐近检验和置信区间

本节是对第 14 章内容的前瞻.假设 Θ 是 $\mathbb{R}^p$ 的一个开子集.定义对数似然比

$$Z(\mathbf{X}, \theta) := 2 \left\{ \log L_{\mathbf{X}}(\hat{\theta}) - \log L_{\mathbf{X}}(\theta) \right\}$$

注意 $Z(\mathbf{X}, \theta) \geq 0$，因为 $\hat{\theta}$ 是（对数）似然函数的最大化点.我们将在第 14 章看到，在某些正则性条件下，

$$Z(\mathbf{X}, \theta) \xrightarrow{\mathcal{D}_\theta} \chi^2_p, \forall \theta$$

这里，“$\xrightarrow{\mathcal{D}_\theta}$”表示在 $\mathbb{P}_\theta$ 下的分布收敛，χ^2_p 表示具有 p 个自由度的卡方分布.

我们称 $Z(\mathbf{X}, \theta)$ 是一个渐近枢轴量：其渐近分布不依赖于未知参数 θ.

对于零假设 $H_0 : \theta = \theta_0$，

一个渐近显著性水平为 α 的检验是：当 $Z\left(\mathbf{X}, \theta_0\right) > \chi^2_p(1-\alpha)$ 时拒绝 H_0，其中 $\chi^2_p(1-\alpha)$ 是 χ^2_p 分布的 $(1-\alpha)$ 分位数.一个渐近 $(1-\alpha)$ 的 θ 的置信集是

$$\begin{aligned} &\left\{\theta : Z(\mathbf{X}, \theta) \leq \chi^2_p(1-\alpha)\right\} \\ =&\left\{\theta : 2 \log L_{\mathbf{X}}(\hat{\theta}) \leq 2 \log L_{\mathbf{X}}(\theta) + \chi^2_p(1-\alpha)\right\} \end{aligned}$$

例子 2.5.1 正态分布的似然比

以下是一个简单的例子.令 X 服从 $\mathcal{N}(\mu, 1)$ 分布，其中 $\mu \in \mathbb{R}$ 为未知参数. μ 的 MLE 是样本均值 $\hat{\mu} = \bar{X}$.

有

$$\log L_{\mathbf{X}}(\hat{\mu}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

并且

$$2 \left\{ \log L_{\mathbf{X}}(\hat{\mu}) - \log L_{\mathbf{X}}(\mu) \right\} = n(\bar{X} - \mu)^2$$

随机变量 $\sqrt{n}(\bar{X}-\mu)$ 在 $\mathbb{P}_\mu$ 下服从 $\mathcal{N}(0,1)$ 分布.因此，其平方 $n(\bar{X}-\mu)^2$ 服从 χ_1^2 分布.因此，在这种情况下，上述检验（或置信区间）是精确的.

第三章插曲：分布理论

3.1 条件分布

回忆条件概率的定义：对于两个集合 A 和 B，若 $P(B)\neq 0$，则给定 B 的条件下 A 的条件概率定义为：

$$P(A\mid B):=\frac{P(A\cap B)}{P(B)}$$

由此可得：

$$P(B\mid A)=P(A\mid B)\frac{P(B)}{P(A)}$$

并且，对于一个分割 $\{B_j\}$[1]，有$P(A)=\sum_j P\left(A\mid B_j\right)P\left(B_j\right)$.

现在考虑两个随机向量 $X\in\mathbb{R}^n$ 和 $Y\in\mathbb{R}^m$.设 (X,Y) 的密度为 $f_{X,Y}(\cdot,\cdot)$（假设存在相对于 Lebesgue 测度的密度）.

X 的边缘密度为：$f_X(\cdot)=\int f_{X,Y}(\cdot,y)dy$

Y 的边缘密度为：$f_Y(\cdot)=\int f_{X,Y}(x,\cdot)dx$

定义 3.1.1

给定 $Y=y$，X 的条件密度为：

$$f_X(x\mid y):=\frac{f_{X,Y}(x,y)}{f_Y(y)},x\in\mathbb{R}^n$$

[1]$\{B_j\}$ 是一个分割，若对于所有 $j\neq k$ 有 $B_j\cap B_k=\emptyset$，且 $P\left(\cup_j B_j\right)=1$.

因此，有：

$$f_Y(y\mid x)=f_X(x\mid y)\frac{f_Y(y)}{f_X(x)},(x,y)\in\mathbb{R}^{n+m}$$

以及：

$$f_X(x)=\int f_X(x\mid y)f_Y(y)dy,x\in\mathbb{R}^n$$

定义 3.1.2

设 $g:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ 是某个函数.给定 $Y=y$，$g(X,Y)$ 的条件期望定义为：

$$E[g(X,Y) \mid Y=y] := \int f_X(x \mid y) g(x,y) dx$$

因此注意到：

$$E\left[g_1(X)g_2(Y) \mid Y=y\right] = g_2(y)E\left[g_1(X) \mid Y=y\right]$$

记号 我们定义随机变量 $E[g(X,Y) \mid Y]$ 为：

$$E[g(X,Y) \mid Y] := h(Y)$$

其中 $h(y)$ 是函数 $h(y) := E[g(X,Y) \mid Y=y]$.

> 引理 3.1.1 迭代期望引理
> *有：*
>
> $$E[E[g(X,Y) \mid Y]] = Eg(X,Y)$$

证明

定义：

$h(y) := E[g(X,Y) \mid Y=y]$

则：

$$\begin{aligned} Eh(Y) &= \int h(y) f_Y(y) dy = \int E[g(X,Y) \mid Y=y] f_Y(y) dy \\ &= \iint g(x,y) f_{X,Y}(x,y) dx dy = Eg(X,Y) \end{aligned}$$

3.2 多项分布

在一项调查中，人们被询问对某一政治问题的看法.令 X 为赞成的回答数，Y 为反对的回答数，Z 为不确定的回答数.调查总人数为 $n = X+Y+Z$.我们将投票视为有放回的样本，其中 $p_1 = P$（赞成），$p_2 = P$（反对），$p_3 = P$（不确定），满足 $p_1+p_2+p_3 = 1$.则：

$$P(X=x, Y=y, Z=z) = \binom{n}{xyz} p_1^x p_2^y p_3^z, (x,y,z) \in \{0,\dots,n\}, x+y+z=n$$

其中：

$$\binom{n}{xy} := \frac{n!}{x!y!z!}$$

称为多项式系数.

> 引理 3.2.1

X 的边缘分布是 $\text{Binomial}(n, p_1)$ *分布.*

证明 对于 $x \in 0, \ldots, n$，有：

$$\begin{aligned} P(X=x) &= \sum_{y=0}^{n-x} P(X=x, Y=y, Z=n-x-y) \\ &= \sum_{y=0}^{n-x} \binom{n}{xy\\ x-x-y} p_1^x p_2^y (1-p_1-p_2)^{n-x-y} \\ &= \binom{n}{x} p_1^x \sum_{y=0}^{n-x} \binom{n-x}{y} p_2^y (1-p_1-p_2)^{n-x-y} \\ &= \binom{n}{x} p_1^x (1-p_1)^{n-x}. \end{aligned}$$

定义 3.2.1

随机向量 $(N_1, \ldots, N_k)$ 具有参数为 n 和 $p_1, \ldots, p_k$（且 $\sum_{j=1}^k p_j = 1$）的多项分布，则对于所有 $(n_1, \ldots, n_k) \in 0, \ldots, n^k$ 且 $n_1 + \cdots + n_k = n$，有：

$$P(N_1 = n_1, \ldots, N_k = n_k) = \binom{n}{n_1 n_k\\ \cdots} p_1^{n_1} \cdots p_k^{n_k}$$

其中：$\binom{n}{n_1 n_k\\ \cdots} := \frac{n!}{n_1! \cdots n_k!}$.

例子 3.2.1 直方图

设 $X_1, \ldots, X_n$ 是独立同分布的随机变量，$X \in \mathbb{R}$，其分布为 F.

设 $-\infty = a_0 < a_1 < \cdots < a_{k-1} < a_k = \infty$.定义对于 $j = 1, \ldots, k$：

$$\begin{aligned} p_j &:= P(X \in (a_{j-1}, a_j]) = F(a_j) - F(a_{j-1}) \\ \frac{N_j}{n} &:= \frac{\#\{X_i \in (a_{j-1}, a_j]\}}{n} = \hat{F}_n(a_j) - \hat{F}_n(a_{j-1}) \end{aligned}$$

则 $(N_1, \ldots, N_k)$ 服从 $\text{Multinomial}(n, p_1, \ldots, p_k)$ 分布.

3.3 泊松分布

定义 3.3.1

若随机变量 $X \in 0, 1, \ldots$ 具有参数 $\lambda > 0$ 的泊松分布，则对于所有 $x \in 0, 1, \ldots$，有

$$P(X=x)=\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}$$

（参见例子 1.4.1）.

引理 3.3.1

假设 X 和 Y 独立，且 X 服从 $\mathrm{Poisson}(\lambda)$ 分布，Y 服从 $\mathrm{Poisson}(\mu)$ 分布.则 $Z:=X+Y$ 服从 $\mathrm{Poisson}(\lambda+\mu)$ 分布.

证明 对于所有 $z\in 0,1,\ldots$，有

$$\begin{aligned}
P(Z=z)&=\sum_{x=0}^{z}P(X=x,Y=z-x)\\
&=\sum_{x=0}^{z}P(X=x)P(Y=z-x)\\
&=\sum_{x=0}^{z}\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}\mathrm{e}^{-\mu}\frac{\mu^{z-x}}{(z-x)!}\\
&=\mathrm{e}^{-(\lambda+\mu)}\frac{1}{z!}\sum_{x=0}^{z}\binom{z}{x}\lambda^x\mu^{z-x}\\
&=\mathrm{e}^{-(\lambda+\mu)}\frac{(\lambda+\mu)^z}{z!}
\end{aligned}$$

引理 3.3.2

设 $X_1,\ldots,X_n$ 是独立的，并且（对于 $i=1,\ldots,n$），X_i 服从 $\mathrm{Poisson}\,(\lambda_i)$ 分布.定义 $Z:=\sum_{i=1}^{n}X_i$.令 $z\in 0,1,\ldots$.则给定 $Z=z$ 时，$(X_1,\ldots,X_n)$ 的条件分布是参数为 z 和 $p_1,\ldots,p_n$ 的多项分布，其中

$$p_j=\frac{\lambda_j}{\sum_{i=1}^{n}\lambda_i},j=1,\ldots,n$$

证明 首先注意到 Z 服从 $\mathrm{Poisson}\,(\lambda_+)$ 分布，其中 $\lambda_+:=\sum_{i=1}^{n}\lambda_i$.因此，对于所有 $(x_1,\ldots,x_n)\in 0,1,\ldots,z^n$，满足 $\sum_{i=1}^{n}x_i=z$，有：

$$\begin{aligned}
P\left(X_1=x_1,\ldots,X_n=x_n\mid Z=z\right)&=\frac{P\left(X_1=x_1,\ldots,X_n=x_n\right)}{P(Z=z)}\\
&=\frac{\prod_{i=1}^{n}\left(\mathrm{e}^{-\lambda_i}\lambda_i^{x_i}/x_i!\right)}{\mathrm{e}^{-\lambda_+}\lambda_+^{z}/z!}\\
&=\binom{z}{x_1\cdots x_n}\left(\frac{\lambda_1}{\lambda_+}\right)^{x_1}\cdots\left(\frac{\lambda_n}{\lambda_+}\right)^{x_n}.
\end{aligned}$$

3.4 两个随机变量最大值的分布

设 X_1 和 X_2 独立且均服从分布 F.假设 F 对 Lebesgue 测度具有密度 f.

定义：

$Z := \max\{X_1, X_2\}$

引理 3.4.1

Z 的分布函数为 F^2.此外，Z 的密度为：

$$f_Z(z) = 2F(z)f(z), z \in \mathbb{R}$$

证明 对于所有 z，有：

$$\begin{aligned} P(Z \leq z) &= P\left(\max\{X_1, X_2\} \leq z\right) \\ &= P\left(X_1 \leq z, X_2 \leq z\right) = F^2(z) \end{aligned}$$

若 F 具有密度 f，则（Lebesgue）几乎处处，$f(z) = \frac{d}{dz}F(z)$.

因此，F^2 的导数几乎处处存在，且 $\frac{d}{dz}F^2(z) = 2F(z)f(z)$

$Z = z$ 时，X_1 的条件分布函数为：

$$F_{X_1}\left(x_1 \mid z\right) = \begin{cases} \frac{F(x_1)}{2F(z)}, & x_1 < z \\ 1, & x_1 \geq z \end{cases}$$

因此注意到，该分布在 z 处有大小为 $\frac{1}{2}$ 的跳跃.

第 4章 完备性与指数族

在本章中，我们将数据记为 $X \in \mathcal{X}$.（在具体例子中，X 通常被替换为 $\mathbf{X} = (X_1, \ldots, X_n)$，其中 $X_1, \ldots, X_n$ 是 X 的独立同分布样本.）我们假设 X 的分布为 $P \in \{P_\theta : \theta \in \Theta\}$.

4.1 完备性

设 $S : \mathcal{X} \to \mathcal{Y}$ 是一个给定的映射.我们考虑统计量 $S = S(X)$.在整个讨论中，我们都会经常用到一个用语:“对于所有可能的 s”,这表示对于定义了 $S = s$ 的条件分布的所有 s（换句话说，对于 S 的分布的支持中所有可能的 s，当然其支持可能依赖于 θ）.

定义 4.1.1

如果对于所有 θ 和所有可能的 s，条件分布

$$P_\theta(X \in \cdot \mid S(X) = s)$$

不依赖于 θ，我们称 S 对于 $\theta \in \Theta$ 是完备的.

例子 4.1.1 完备性与伯努利试验

设 $\mathbf{X}=(X_1,\ldots,X_n)$，其中 $X_1,\ldots,X_n$ 独立同分布，服从伯努利分布，成功概率为 $\theta\in(0,1)$：对于 $i=1,\ldots,n$，

$$P_\theta\left(X_i=1\right)=1-P_\theta\left(X_i=0\right)=\theta$$

取 $S=\sum_{i=1}^n X_i$.则 S 对于 θ 是完备的：
对于所有可能的 s，

$$\mathbb{P}_\theta\left(X_1=x_1,\ldots,X_n=x_n\mid S=s\right)=\frac{1}{\binom{n}{s}},\sum_{i=1}^n x_i=s$$

例子 4.1.2 完备性与泊松分布

设 $\mathbf{X}:=(X_1,\ldots,X_n)$，其中 $X_1,\ldots,X_n$ 独立同分布且服从泊松分布 $\mathrm{Poisson}(\theta)$.取 $S=\sum_{i=1}^n X_i$.则 S 服从泊松分布 $\mathrm{Poisson}(n\theta)$.对于所有可能的 s，$\mathbf{X}$ 在给定 $S=s$ 条件下的条件分布为参数为 s 和 $(p_1,\ldots,p_n)=\left(\frac{1}{n},\ldots,\frac{1}{n}\right)$ 的多项分布：

$$\mathbb{P}_\theta\left(X_1=x_1,\ldots,X_n=x_n\mid S=s\right)=\binom{s}{x_1,\ldots,x_n}\prod_{i=1}^n p_i^{x_i}$$

由于此分布不依赖于 θ，所以 S 对于 θ 是完备的.

例子 4.1.3 完备性与指数分布

设 X_1 和 X_2 是独立的，且都服从参数为 $\theta>0$ 的指数分布.
例如，X_1 的密度为：

$$f_{X_1}(x;\theta)=\theta\mathrm{e}^{-\theta x},x>0$$

取 $S=X_1+X_2$.可以验证 S 的密度为：

$$f_S(s;\theta)=s\theta^2\mathrm{e}^{-\theta s},s>0$$

（这是 $\mathrm{Gamma}(2,\theta)$ 分布.）对于所有可能的 s，(X_1,X_2) 在给定 $S=s$ 条件下的条件密度为：$f_{X_1,X_2}\left(x_1,x_2\mid S=s\right)=\frac{1}{s},x_1+x_2=s.$

因此，S 对于 θ 是完备的.

例子 4.1.4 顺序统计量的完备性

设 $X_1,\ldots,X_n$ 是来自连续分布 F 的独立同分布样本.则 $S:=\left(X_{(1)},\ldots,X_{(n)}\right)$ 对于 F 是完备的：对于所有可能的 $s=(s_1,\ldots,s_n)$（$s_1<\ldots<s_n$），以及 $(x_1,\ldots,x_n)=s$，

$$\mathbb{P}_\theta\left((X_1,\ldots,X_n)=(x_1,\ldots,x_n)\mid\left(X_{(1)},\ldots,X_{(n)}\right)=s\right)=\frac{1}{n!}$$

例子 4.1.5 完备性与均匀分布

设 X_1 和 X_2 独立，且均服从区间 $[0,\theta]$ 上的均匀分布，其中 $\theta>0$.定义 $Z:=X_1+X_2$.

引理 随机变量 Z 的密度为：

f_{Z}(z ; \theta)=\left\{\begin{array}{ll} z / \theta^{2} & \text { if } 0 \leq z \le

证明 首先，假设 $\theta=1$.则 Z 的分布函数为：

FZ(z)={z2/20≤z≤11−(2−z)2/21≤z≤2F_{Z}(z)=\begin{cases}z^{2} / 2 & 0 \leq z \leq 1 \ 1-(2-z)^{2} / 2 & 1 \leq z \leq 2\end{cases}

因此，密度为：

fZ(z)={z0≤z≤12−z1≤z≤2.f_{Z}(z)=\left\{\begin{array}{ll} z & 0 \leq z \leq 1 \ 2-z & 1 \leq z \leq 2 \end{array} .\right.

对于一般的 θ，结果由变换 $Z \mapsto \theta Z$ 得到，此变换将 f_Z 映射为 $f_Z(\cdot/\theta)/\theta$.

随机变量 (X_1, X_2) 在给定 $Z=z \in (0, 2\theta)$ 条件下的条件密度依赖于 θ，因此 Z 对于 θ 不是完备的.

现在考虑 $S:=\max\{X_1, X_2\}$.随机变量 (X_1, X_2) 在给定 $S=s \in (0, \theta)$ 条件下的条件分布不依赖于 θ（这里不做证明），因此 S 对于 θ 是完备的.

4.2 内曼分解定理

定理 4.2.1（内曼分解定理）
假设 $\{P_\theta : \theta \in \Theta\}$ 被某个 σ-有限测度 ν 主导.记 $p_\theta := dP_\theta/d\nu$ 为密度函数.那么，当且仅当 p_θ 可以写成以下形式时，统计量 S 对于 θ 是充分的：

$$p_\theta(x) = g_\theta(S(x))h(x), \forall x, \theta,$$

其中 $g_\theta(\cdot) \geq 0$ 和 $h(\cdot) \geq 0$ 是某些函数.

证明（离散情形）
假设 X 只取值于 $a_1, a_2, \ldots \forall \theta$（因此我们可以取 ν 为计数测度）.令 Q_θ 为 S 的分布：

$$Q_\theta(s) := \sum_{j:S(a_j)=s} P_\theta\left(X=a_j\right)$$

S 给定的条件下 X 的条件分布为：

$$P_\theta(X=x \mid S=s)=\frac{P_\theta(X=x)}{Q_\theta(s)}, \quad S(x)=s.$$

($\Rightarrow$) 充分性推出分解形式
若 S 对 θ 充分，上述条件分布不依赖于 θ，仅仅是 x 的一个函数，记为 $h(x)$.因此，对于 $S(x)=s$，我们可以写为：

$$P_\theta(X=x)=P_\theta(X=x \mid S=s) Q_\theta(S=s)=h(x) g_\theta(s),$$

其中 $g_\theta(s)=Q_\theta(S=s)$.

($\Leftarrow$) 分解形式推出充分性
将 $p_\theta(x) = g_\theta(S(x))h(x)$ 代入，可以得到：

$$Q_\theta(s) = g_\theta(s) \sum_{j:S(a_j)=s} h\left(a_j\right).$$

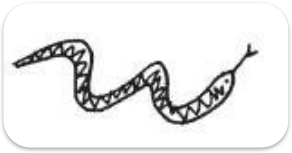
将其代入 $P_\theta(X = x \mid S = s)$ 的公式中：

$$P_\theta(X = x \mid S = s) = \frac{h(x)}{\sum_{j:S(a_j)=s} h\left(a_j\right)},$$

可以看到这不依赖于 θ.

备注

对于一般情形的证明与此类似，但存在一些微妙之处！



例子 4.2.1 不确定端点的均匀分布的充分性

令 $X_1, \ldots, X_n$ 为独立同分布，且在区间 $[0, \theta]$ 上均匀分布.则 $\mathbf{X} = (X_1, \ldots, X_n)$ 的密度为：

$$\begin{aligned}\mathbf{p}_\theta\left(x_1, \ldots, x_n\right) &= \frac{1}{\theta^n} 1\left\{0 \leq \min\left\{x_1, \ldots, x_n\right\} \leq \max\left\{x_1, \ldots, x_n\right\} \leq \theta\right\} \\ &= g_\theta\left(S\left(x_1, \ldots, x_n\right)\right) h\left(x_1, \ldots, x_n\right),\end{aligned}$$

其中：

$$g_\theta(s) := \frac{1}{\theta^n} 1\{s \leq \theta\},$$

以及：

$$h\left(x_1, \ldots, x_n\right) := 1\left\{0 \leq \min\left\{x_1, \ldots, x_n\right\}\right\}.$$

因此，$S = \max\left\{X_1, \ldots, X_n\right\}$ 对 θ 是充分的.

推论 4.2.1

似然函数为 $L_X(\theta) = p_\theta(X) = g_\theta(S)h(X)$.因此，最大似然估计
$\hat{\theta} = \arg\max_\theta L_X(\theta) = \arg\max_\theta g_\theta(S)$ 仅依赖于充分统计量 S.

4.3 指数族

定义 4.3.1

一个 k 维指数族是一族分布 $\{P_\theta : \theta \in \Theta\}$，被某个 σ -有限测度 ν 主导，其密度 $p_\theta = dP_\theta/d\nu$ 具有以下形式：

$$p_\theta(x) = \exp\left[\sum_{j=1}^{k} c_j(\theta) T_j(x) - d(\theta)\right] h(x),$$

其中：

- $T_j(x)$ 是统计量；
- $c_j(\theta)$ 是参数的函数；
- $d(\theta)$ 和 $h(x)$ 分别与参数 θ 和样本 x 有关.

注解

1. 在 k 维指数族的情况下，k 维统计量 $S(X) = (T_1(X), \ldots, T_k(X))$ 是 θ 的充分统计量.
2. 如果 $X_1, \ldots, X_n$ 是从一个 k 维指数族中独立同分布的样本，那么 $\mathbf{X} = (X_1, \ldots, X_n)$ 的分布也属于 k 维指数族，其密度为：

   $$\mathbf{p}_\theta(\mathbf{x}) = \prod_{i=1}^{n} p_\theta\left(x_i\right) = \exp\left[\sum_{j=1}^{k} n c_j(\theta) \bar{T}_j(\mathbf{x}) - n d(\theta)\right] \prod_{i=1}^{n} h\left(x_i\right),$$

   其中，对于 $j = 1, \ldots, k$，

   $$\bar{T}_j(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} T_j\left(x_i\right).$$

   因此，$S(\mathbf{X}) = \left(\bar{T}_1(\mathbf{X}), \ldots, \bar{T}_k(\mathbf{X})\right)$ 对 θ 是充分的.
3. 函数 $\{T_j\}$ 和 $\{c_j\}$ 并非唯一确定.

例子 4.3.1 泊松分布

若 X 服从泊松分布 $\text{Poisson}(\theta)$，则其密度为：

$$\begin{aligned} p_\theta(x) &= e^{-\theta} \frac{\theta^x}{x!} \\ &= \exp[x \log \theta - \theta] \frac{1}{x!}. \end{aligned}$$

因此，我们可以取 $T(x) = x, c(\theta) = \log \theta, d(\theta) = \theta.$

例子 4.3.2 二项分布

若 X 服从二项分布 $\text{Binomial}(n, \theta)$，则其密度为：

$$
\begin{aligned}
p_\theta(x) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\
&= \binom{n}{x}\left(\frac{\theta}{1-\theta}\right)^x(1-\theta)^n \\
&= \binom{n}{x}\exp\left[x\log\left(\frac{\theta}{1-\theta}\right)+n\log(1-\theta)\right].
\end{aligned}
$$

因此，我们可以取 $T(x)=x, c(\theta)=\log(\theta/(1-\theta)), d(\theta)=-n\log(1-\theta)$.

例子 4.3.3 负二项分布

若 X 服从负二项分布 $\text{NegBinomial}(m,\theta)$，其中 m 已知，则其密度为：

$$
\begin{aligned}
p_\theta(x) &= \frac{\Gamma(x+m)}{\Gamma(m)x!}\theta^m(1-\theta)^x \\
&= \frac{\Gamma(x+m)}{\Gamma(m)x!}\exp[x\log(1-\theta)+m\log(\theta)].
\end{aligned}
$$

因此，我们可以取 $T(x)=x, c(\theta)=\log(1-\theta), d(\theta)=-m\log(\theta)$.

例子 4.3.4 一参数伽马分布

若 X 服从 $\text{Gamma}(m,\theta)$ 分布，其中 m 已知，则其密度为：

$$
\begin{aligned}
p_\theta(x) &= \mathrm{e}^{-\theta x}x^{m-1}\frac{\theta^m}{\Gamma(m)} \\
&= \frac{x^{m-1}}{\Gamma(m)}\exp[-\theta x+m\log\theta].
\end{aligned}
$$

因此，我们可以取 $T(x)=x, c(\theta)=-\theta, d(\theta)=-m\log\theta$.

例子 4.3.5 两参数伽马分布

若 X 服从 $\text{Gamma}(m,\lambda)$ 分布，令 $\theta=(m,\lambda)$，则其密度为：

$$
\begin{aligned}
p_\theta(x) &= \mathrm{e}^{-\lambda x}x^{m-1}\frac{\lambda^m}{\Gamma(m)} \\
&= \exp[-\lambda x+(m-1)\log x+m\log\lambda-\log\Gamma(m)].
\end{aligned}
$$

因此，我们可以取
$T_1(x)=x, T_2(x)=\log x, c_1(\theta)=-\lambda, c_2(\theta)=(m-1), d(\theta)=-m\log\lambda+\log\Gamma(m)$.

例子 4.3.6 正态分布

若 X 服从 $\mathcal{N}\left(\mu, \sigma^2\right)$ 分布，令 $\theta=(\mu, \sigma)$，则其密度为：

$$\begin{aligned} p_\theta(x) & =\frac{1}{\sqrt{2 \pi} \sigma} \exp \left[-\frac{(x-\mu)^2}{2 \sigma^2}\right] \\ & =\frac{1}{\sqrt{2 \pi}} \exp \left[\frac{x \mu}{\sigma^2}-\frac{x^2}{2 \sigma^2}-\frac{\mu^2}{2 \sigma^2}-\log \sigma\right] . \end{aligned}$$

因此，我们可以取
$T_1(x)=x, T_2(x)=x^2, c_1(\theta)=\mu / \sigma^2, c_2(\theta)=-1 / \left(2 \sigma^2\right), d(\theta)=\mu^2 / \left(2 \sigma^2\right)+\log (\sigma)$.

4.4 插曲：随机向量的均值和协方差矩阵

设 $Z \in \mathbb{R}^k$ 是一个随机向量，那么其均值（如果存在）定义为由 Z 每个分量的均值组成的向量：

$$EZ = \begin{pmatrix} EZ_1 \\ \vdots \\ EZ_k \end{pmatrix}$$

Z 的协方差矩阵（如果存在）定义为对称的 $k \times k$ 矩阵 Σ，其元素表示 Z 中每一对分量的协方差：

$$\Sigma := EZZ' - EZEZ' = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,k} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,k} & \sigma_{2,k} & \cdots & \sigma_k^2 \end{pmatrix}$$

其中：

- Z' 表示 Z 的转置；
- $\sigma_j^2 := \operatorname{var}(Z_j)$ 表示 Z_j 的方差；
- $\sigma_{j_1,j_2} = \operatorname{cov}(Z_{j_1}, Z_{j_2})$ 表示 Z_{j_1} 与 Z_{j_2} 的协方差.

协方差矩阵 Σ 通常记为 $\operatorname{Cov}(Z)$.

4.5 指数族的标准形式

在这一节中，我们假设正则性条件（如导数和逆的存在，以及可交换微分与积分次序）.

设 $\Theta \subset \mathbb{R}^k$，$\{P_\theta : \theta \in \Theta\}$ 是由某 σ -有限测度 ν 主导的一族概率测度.定义其密度为：

$$p_\theta := \frac{dP_\theta}{d\nu}$$

定义

我们称 $\{P_\theta : \theta \in \Theta\}$ 为标准形式的指数族，如果其密度可以写为：

$$p_\theta(x) = \exp\left[\sum_{j=1}^{k} \theta_j T_j(x) - d(\theta)\right] h(x)$$

其中，$d(\theta)$ 是归一化常数：

$$d(\theta) = \log\left(\int \exp\left[\sum_{j=1}^{k} \theta_j T_j(x)\right] h(x) d\nu(x)\right)$$

我们定义：

- $d(\theta)$ 的一阶导数为

$$\dot{d}(\theta) := \frac{\partial}{\partial \theta} d(\theta)$$

- $d(\theta)$ 的二阶导数矩阵为

$$\ddot{d}(\theta) := \frac{\partial^2}{\partial \theta \partial \theta'} d(\theta) = \left(\frac{\partial^2}{\partial \theta_{j_1} \partial \theta_{j_2}} d(\theta)\right)$$

我们进一步记：

- $T(X)$ 为

$$T(X) := \begin{pmatrix} T_1(X) \\ \vdots \\ T_k(X) \end{pmatrix}$$

- $\mathbb{E}_\theta T(X)$ 为

$$\mathbb{E}_\theta T(X) := \begin{pmatrix} \mathbb{E}_\theta T_1(X) \\ \vdots \\ \mathbb{E}_\theta T_k(X) \end{pmatrix}$$

- $T(X)$ 的协方差矩阵为

$$\mathrm{Cov}_\theta(T(X)) := \mathbb{E}_\theta T(X) T'(X) - \mathbb{E}_\theta T(X) \mathbb{E}_\theta T'(X)$$

引理 4.5.1

在正则性条件下：

$$\mathbb{E}_\theta T(X) = \dot{d}(\theta), \mathrm{Cov}_\theta(T(X)) = \ddot{d}(\theta)$$

证明

由 $d(\theta)$ 的定义，有：

$$\begin{aligned}\dot{d}(\theta) &= \frac{\partial}{\partial\theta}\log\left(\int \exp\left[\theta' T(x)\right]h(x)d\nu(x)\right)\\ &= \frac{\int \exp\left[\theta' T(x)\right]T(x)h(x)d\nu(x)}{\int \exp\left[\theta' T(x)\right]h(x)d\nu(x)}\\ &= \int \exp\left[\theta' T(x) - d(\theta)\right]T(x)h(x)d\nu(x)\\ &= \int p_\theta(x)T(x)d\nu(x) = \mathbb{E}_\theta T(X)\end{aligned}$$

类似地，对 $d(\theta)$ 的二阶导数：

$$\begin{aligned}\ddot{d}(\theta) &= \int TT' p_\theta d\nu - \left(\int p_\theta T d\nu\right)\left(\int p_\theta T' d\nu\right)\\ &= \mathbb{E}_\theta T(X)T'(X) - \left(\mathbb{E}_\theta T(X)\right)\left(\mathbb{E}_\theta T'(X)\right)\\ &= \mathrm{Cov}_\theta(T(X))\end{aligned}$$

4.6 一维情况的重新参数化

我们现在简化为一维情况，即 $\Theta \subset \mathbb{R}$.考虑一个（不一定是标准形式的）指数族：

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x).$$

通过重新参数化

$$\theta \mapsto c(\theta) := \gamma \text{ (记为}\gamma\text{)},$$

可以将其改写为标准形式：

$$\tilde{p}_\gamma(x) = \exp\left[\gamma T(x) - d_0(\gamma)\right]h(x),$$

其中当 c 是一一映射时，有

$$d_0(\gamma) = d\left(c^{-1}(\gamma)\right)$$

因此，

$$\mathbb{E}_\theta T(X) = \dot{d}_0(\gamma) = \frac{\dot{d}\left(c^{-1}(\gamma)\right)}{\dot{c}\left(c^{-1}(\gamma)\right)} = \frac{\dot{d}(\theta)}{\dot{c}(\theta)} \tag{4.1}$$

而

$$\begin{aligned}\mathrm{var}_\theta(T(X)) = \ddot{d}_0(\gamma) &= \frac{\ddot{d}\left(c^{-1}(\gamma)\right)}{\left[\dot{c}\left(c^{-1}(\gamma)\right)\right]^2} - \frac{\dot{d}\left(c^{-1}(\gamma)\right)\ddot{c}\left(c^{-1}(\gamma)\right)}{\left[\dot{c}\left(c^{-1}(\gamma)\right)\right]^3}\\ &= \frac{\ddot{d}(\theta)}{[\dot{c}(\theta)]^2} - \frac{\dot{d}(\theta)\ddot{c}(\theta)}{[\dot{c}(\theta)]^3}\\ &= \frac{1}{[\dot{c}(\theta)]^2}\left(\ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)}\ddot{c}(\theta)\right).\end{aligned}$$

4.7 Score 函数和 Fisher 信息

考虑一个任意的（但满足正则性条件的）密度族 $\{p_\theta : \theta \in \Theta\}$，且为简化 $\Theta \subset \mathbb{R}$.

定义 4.7.1 Score 函数

$$s_\theta(x) := \frac{d}{d\theta} \log p_\theta(x).$$

估计 θ 的 Fisher 信息为：

$$I(\theta) := \operatorname{var}_\theta \left(s_\theta(X) \right)$$

更一般地，估计参数 θ 的可微函数 $g(\theta)$ 的 Fisher 信息为 $I(\theta)/[\dot{g}(\theta)]^2$.

见第 5 章和第 13 章.

引理 4.7.1

在正则性条件下，有

$$\mathbb{E}_\theta s_\theta(X) = 0$$

且

$$I(\theta) = -\mathbb{E}_\theta \dot{s}_\theta(X)$$

其中 $\dot{s}_\theta(x) := \frac{d}{d\theta} s_\theta(x)$.

证明

结果来源于密度函数的积分为 1 这一事实，并假设可以交换微分和积分的顺序：

$$\begin{aligned}
\mathbb{E}_\theta s_\theta(X) &= \int s_\theta(x) p_\theta(x) d\nu(x) \\
&= \int \frac{d \log p_\theta(x)}{d\theta} p_\theta(x) d\nu(x) = \int \frac{\dot{p}_\theta(x)}{p_\theta(x)} p_\theta(x) d\nu(x) \\
&= \int \dot{p}_\theta(x) d\nu(x) = \frac{d}{d\theta} \int p_\theta(x) d\nu(x) = \frac{d}{d\theta} 1 = 0,
\end{aligned}$$

且

$$\begin{aligned}
\mathbb{E}_\theta \dot{s}_\theta(X) &= \mathbb{E}_\theta \left[\frac{\ddot{p}_\theta(X)}{p_\theta(X)} - \left(\frac{\dot{p}_\theta(X)}{p_\theta(X)} \right)^2 \right] \\
&= \mathbb{E}_\theta \left[\frac{\ddot{p}_\theta(X)}{p_\theta(X)} \right] - \mathbb{E}_\theta s_\theta^2(X)
\end{aligned}$$

其中，$\mathbb{E}_\theta s_\theta^2(X)$ 等于 $\operatorname{var}_\theta s_\theta(X)$，因为 $\mathbb{E}_\theta s_\theta(X) = 0$.进一步有：

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{\ddot{p}_\theta(X)}{p_\theta(X)}\right] &= \int \frac{d^2}{d\theta^2} p_\theta(x) d\nu(x) \\ &= \frac{d^2}{d\theta^2}\int p_\theta(x) d\nu(x) \\ &= \frac{d^2}{d\theta^2} 1 = 0.\end{aligned}$$

4.8 指数族的 Score 函数

在一维指数族的特殊情况下，概率分布 $\{P_\theta : \theta \in \Theta\}$ 的密度为

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$$

因此，

$$s_\theta(x) = \dot{c}(\theta)T(x) - \dot{d}(\theta)$$

由于 $\mathbb{E}_\theta s_\theta(X) = 0$，我们得到

$$\mathbb{E}_\theta T(X) = \frac{\dot{d}(\theta)}{\dot{c}(\theta)}$$

这再次验证了 (4.1).此外，

$$\dot{s}_\theta(x) = \ddot{c}(\theta)T(x) - \ddot{d}(\theta)$$

因此，利用不等式 $\operatorname{var}_\theta\left(s_\theta(X)\right) = -\mathbb{E}_\theta \dot{s}_\theta(X)$，可以推导出

$$\begin{aligned}\left[\dot{c}(\theta)\right]^2 \operatorname{var}_\theta(T(X)) &= -\ddot{c}(\theta)\mathbb{E}_\theta T(X) + \ddot{d}(\theta) \\ &= \ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)}\ddot{c}(\theta)\end{aligned}$$

这再次验证了 (4.2).进一步得到

$$I(\theta) = \ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)}\ddot{c}(\theta)$$

估计 $\gamma = c(\theta)$ 的 Fisher 信息为

$$I_0(\gamma) = \ddot{d}_0(\gamma) = \frac{I(\theta)}{[\dot{c}(\theta)]^2}$$

例子 4.8.1 伯努利分布的标准形式

令 $X \in \{0, 1\}$ 服从伯努利分布，其成功概率为 $\theta \in (0, 1)$：

$$p_\theta(x) = \theta^x(1-\theta)^{1-x} = \exp\left[x\log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right], x \in \{0, 1\}$$

重新参数化为

$$\gamma := c(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$$

这被称为对数赔率比 (log-odds ratio).反转后得到

$$\theta = \frac{\mathrm{e}^{\gamma}}{1+\mathrm{e}^{\gamma}}$$

因此，

$$d(\theta) = -\log(1-\theta) = \log\left(1+\mathrm{e}^{\gamma}\right) := d_0(\gamma)$$

因此，

$$\dot{d}_0(\gamma) = \frac{\mathrm{e}^{\gamma}}{1+\mathrm{e}^{\gamma}} = \theta = \mathbb{E}_{\theta}X$$

和

$$\ddot{d}_0(\gamma) = \frac{\mathrm{e}^{\gamma}}{1+\mathrm{e}^{\gamma}} - \frac{\mathrm{e}^{2\gamma}}{\left(1+\mathrm{e}^{\gamma}\right)^2} = \frac{\mathrm{e}^{\gamma}}{\left(1+\mathrm{e}^{\gamma}\right)^2} = \theta(1-\theta) = \mathrm{var}_{\theta}(X)$$

Score 函数为

$$\begin{aligned} s_{\theta}(x) &= \frac{d}{d\theta}\left[x\log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right] \\ &= \frac{x}{\theta(1-\theta)} - \frac{1}{1-\theta}. \end{aligned}$$

估计成功概率 θ 的 Fisher 信息为

$$\mathbb{E}_{\theta}s_{\theta}^2(X) = \frac{\mathrm{var}_{\theta}(X)}{[\theta(1-\theta)]^2} = \frac{1}{\theta(1-\theta)}$$

而估计对数赔率比 γ 的 Fisher 信息为

$$I_0(\gamma) = \theta(1-\theta)$$

4.9 最小充分性

定义 4.9.1

若对于所有 θ，存在常数 $c(x,\tilde{x})$ 使得

$$L_x(\theta) = L_{\tilde{x}}(\theta)c(x,\tilde{x}),$$

则称似然函数 $L_x(\theta)$ 和 $L_{\tilde{x}}(\theta)$ 在 $(x,\tilde{x})$ 处是成比例的，记为

$$L_x(\theta) \propto L_{\tilde{x}}(\theta).$$

若统计量 S 满足对于所有 x 和 $\tilde{x}$，只要 $L_x(\theta) \propto L_{\tilde{x}}(\theta)$ 就有 $S(x) = S(\tilde{x})$，则称 S 为**最小充分统计量**.

例子 4.9.1 正态分布的最小充分统计量

设 $X_1, \ldots, X_n$ 是独立的 $\mathcal{N}(\theta, 1)$ 分布随机变量，则 $S = \sum_{i=1}^n X_i$ 是 θ 的充分统计量.此外，

$$\log L_{\mathbf{x}}(\theta) = S(\mathbf{x})\theta - \frac{n\theta^2}{2} - \frac{\sum_{i=1}^n x_i^2}{2} - \log(2\pi)/2$$

因此，

$$\log L_{\mathbf{x}}(\theta) - \log L_{\tilde{\mathbf{x}}}(\theta) = (S(\mathbf{x}) - S(\tilde{\mathbf{x}}))\theta - \frac{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n (\tilde{x}_i)^2}{2},$$

这仅在 $S(\mathbf{x}) = S(\tilde{\mathbf{x}})$ 时等于某个 $c(\mathbf{x}, \tilde{\mathbf{x}})$ 的对数.因此，S 是最小充分统计量.

例子 4.9.2 拉普拉斯分布的最小充分统计量

设 $X_1, \ldots, X_n$ 是独立的拉普拉斯分布随机变量，位置参数为 θ.则

$$\log L_{\mathbf{x}}(\theta) = -(\log 2)/2 - \sqrt{2}\sum_{i=1}^n |x_i - \theta|$$

因此，

$$\log L_{\mathbf{x}}(\theta) - \log L_{\tilde{\mathbf{x}}}(\theta) = -\sqrt{2}\sum_{i=1}^n (|x_i - \theta| - |\tilde{x}_i - \theta|)$$

当且仅当 $(x_{(1)}, \ldots, x_{(n)}) = (\tilde{x}_{(1)}, \ldots, \tilde{x}_{(n)})$ 时，上式等于某个 $c(\mathbf{x}, \tilde{\mathbf{x}})$ 的对数.因此，次序统计量 $X_{(1)}, \ldots, X_{(n)}$ 是最小充分统计量.

第 5 章偏差、方差与 Cramér-Rao 下界

5.1 什么是无偏估计?

设 $X \in \mathcal{X}$ 是观测值，其分布 P 属于一组分布 $\{P_\theta : \theta \in \Theta\}$.感兴趣的参数为 $\gamma := g(\theta)$，其中 $g : \Theta \to \mathbb{R}$.除本章最后部分外，参数 γ 假设为一维.

令 $T : \mathcal{X} \to \mathbb{R}$ 为 $g(\theta)$ 的估计量.

定义 5.1.1

$T = T(X)$ 的偏差为

$$\text{bias}_\theta(T) := \mathbb{E}_\theta T - g(\theta)$$

若对所有 θ，均有

$$\text{bias}_\theta(T) = 0,$$

则称 T 为无偏估计量.

因此，无偏性意味着没有系统误差：$\mathbb{E}_\theta T = g(\theta)$.

例子 5.1.1 二项分布中的无偏估计量

设 $X \sim \text{Binomial}(n, \theta)$，其中 $0 < \theta < 1$.则

$$\mathbb{E}_\theta T(X) = \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} T(k) =: q(\theta)$$

注意 $q(\theta)$ 是 θ 的次数不超过 n 的多项式.因此，只有次数不超过 n 的多项式形式的参数 $g(\theta)$ 才能被无偏估计.例如，$\sqrt{\theta}$ 或 $\theta/(1-\theta)$ 不存在无偏估计量.

例子 5.1.2 泊松分布中的无偏估计量

设 $X \sim \text{Poisson}(\theta)$，则

$$\mathbb{E}_\theta T(X) = \sum_{k=0}^{\infty} \mathrm{e}^{-\theta} \frac{\theta^k}{k!} T(k) =: \mathrm{e}^{-\theta} \mathrm{p}(\theta).$$

注意 $\mathrm{p}(\theta)$ 是关于 θ 的幂级数.因此，仅当参数 $g(\theta)$ 是 θ 的幂级数乘以 $\mathrm{e}^{-\theta}$ 时，才可以无偏估计.例如，早期失效的概率

$$g(\theta) := \mathrm{e}^{-\theta} = P_\theta(X = 0)$$

是一个可无偏估计的参数.一个 $\mathrm{e}^{-\theta}$ 的无偏估计量为

$$T(X) = 1\{X = 0\}.$$

另一个例子是当目标参数为

$$g(\theta) := \mathrm{e}^{-2\theta}$$

时，其无偏估计量为

$$T(X) = \begin{cases} +1 & \text{若 } X \text{ 为偶数} \\ -1 & \text{若 } X \text{ 为奇数} \end{cases}$$

然而，这个估计量毫无实际意义！

例子 5.1.3 方差的无偏估计量

设 $X_1, \ldots, X_n$ 是独立同分布的 $\mathcal{N}(\mu, \sigma^2)$ 随机变量，并令 $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$.定义

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

则 S^2 是 σ^2 的无偏估计量.但 S 并不是 σ 的无偏估计量.是否可以构造一个 σ 的无偏估计量?

结论

要求无偏性可能存在以下缺点:

1. 无偏估计量并不总是存在.
2. 即使存在，它们有时可能毫无意义.
3. 无偏性的性质在进行非线性变换时并不保留.

5.2 UMVU（统一最小方差无偏估计量）

定义 5.2.1

估计量 T 的均方误差（Mean Square Error, MSE）定义为:

$$\mathrm{MSE}_\theta(T) := \mathbb{E}_\theta(T - g(\theta))^2$$

引理 5.2.1

均方误差可以分解为:

$$\mathrm{MSE}_\theta(T) = \mathrm{bias}_\theta^2(T) + \mathrm{var}_\theta(T)$$

证明: 设 $\mathbb{E}_\theta T = q(\theta)$，则

$$\begin{aligned}\mathbb{E}_\theta(T - g(\theta))^2 = &\underbrace{\mathbb{E}_\theta(T - q(\theta))^2}_{=\mathrm{var}_\theta(T)} + \underbrace{(q(\theta) - g(\theta))^2}_{=\mathrm{bias}_\theta^2(T)} \\ &+2(q(\theta) - g(\theta))\underbrace{\mathbb{E}_\theta(T - q(\theta))}_{=0}.\end{aligned}$$

这说明均方误差由两部分组成：平方偏差和方差.这种分解称为**偏差-方差分解**.通常，减少偏差会导致方差增加，反之亦然.

例子 5.2.1 正态分布中的估计量

设 $X_1, \ldots, X_n$ 是独立同分布的 $\mathcal{N}(\mu, \sigma^2)$ 随机变量，参数为 $\theta := (\mu, \sigma^2)$.

情形 i：目标是估计均值 μ

定义估计量 $T := a\bar{X}$，其中 $0 \le a \le 1$.此时，偏差随 a 减小，而方差随 a 增大:

$$\mathrm{MSE}_\theta(T) = \mathbb{E}_\theta(T-\mu)^2 = (1-a)^2\mu^2 + a^2\frac{\sigma^2}{n}$$

右侧关于 a 的最小值在以下位置取得：

$$a_{\text{opt}} := \frac{\mu^2}{\frac{\sigma^2}{n} + \mu^2}$$

然而，由于 a_{opt} 依赖于未知参数，因此无法直接用于构造估计量.

情形 ii：目标是估计方差 σ^2

定义样本方差：

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

已知 S^2 是 σ^2 的无偏估计量.我们将其与以下估计量比较：

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

为了计算这两个估计量的均方误差，回顾以下性质：

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

χ^2 分布的期望和方差为：

$$E\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}\right) = n-1, \quad \mathrm{var}\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}\right) = 2(n-1)$$

因此，

$$\mathrm{MSE}_\theta(S^2) = \frac{2\sigma^4}{n-1}$$

而对于 $\hat{\sigma}^2$，

$$\mathbb{E}_\theta\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2, \quad \mathrm{bias}_\theta(\hat{\sigma}^2) = -\frac{1}{n}\sigma^2$$

其均方误差为：

$$\begin{aligned}\mathrm{MSE}_\theta(\hat{\sigma}^2) &= \frac{\sigma^4}{n^2} + \frac{2(n-1)\sigma^4}{n^2} \\ &= \frac{(2n-1)\sigma^4}{n^2}\end{aligned}$$

结论：$\hat{\sigma}^2$ 的均方误差小于 S^2 的均方误差！

定义 5.2.2

若估计量 T^* 满足以下条件，则称其为 UMVU（统一最小方差无偏估计量）：

$$\mathrm{var}_\theta(T^*) \leq \mathrm{var}_\theta(T), \quad \forall \theta$$

设 T 是无偏估计量，且 S 是充分统计量.定义

$$T^* := E(T \mid S)$$

由于 T 的条件分布 $P(T \mid S)$ 不依赖于 θ，因此 T^* 是一个估计量.此外，由迭代期望引理可知：

$$\mathbb{E}_\theta T^* = \mathbb{E}_\theta[E(T \mid S)] = \mathbb{E}_\theta T = g(\theta)$$

通过对 S 的条件期望，"多余"的样本方差被消除.引理表明：

$$\mathrm{var}_\theta(T^*) \leq \mathrm{var}_\theta(T), \quad \forall \theta$$

引理 5.2.2

设 Y 和 Z 是随机变量，则有：

$$\mathrm{var}(Y) = \mathrm{var}(E(Y \mid Z)) + E[\mathrm{var}(Y \mid Z)]$$

证明：

$$\begin{aligned}\mathrm{var}(E(Y \mid Z)) &= E[E(Y \mid Z)^2] - [E(E(Y \mid Z))]^2 \\ &= E[E(Y \mid Z)^2] - [EY]^2\end{aligned}$$

$$E[\mathrm{var}(Y \mid Z)] = E[E(Y^2 \mid Z) - E(Y \mid Z)^2] = EY^2 - E[E(Y \mid Z)^2]$$

加总后，得到方差公式：

$$\mathrm{var}(Y) = EY^2 - [EY]^2$$

5.3 Lehmann-Scheffé 引理

我们需要回答的问题是：是否可以构造一个无偏估计量，其方差比 $T^* = E(T \mid S)$ 更小？注意，T^* 仅依赖于充分统计量 $S = S(X)$，也就是说它仅依赖于充分统计量.在寻找 UMVU（统一最小方差无偏估计量）时，可以将注意力集中在仅依赖于 S 的估计量上.因此，如果仅存在一个仅依赖于 S 的无偏估计量，则它必须是 UMVU.

定义 5.3.1

若充分统计量 S 满足以下条件，则称其为**完全统计量**：

$$\mathbb{E}_\theta h(S) = 0 \quad \forall \theta \implies h(S) = 0,\ P_\theta - \text{a.s.},\ \forall \theta$$

其中，h 是关于 S 的函数，与参数 θ 无关.

引理 5.3.1 (Lehmann-Scheffé 引理)

令 T 是 $g(\theta)$ 的无偏估计量，且对于所有 θ，T 的方差有限.此外，设 S 是充分且完全统计量.则 $T^* := E(T \mid S)$ 是 UMVU.

证明:

我们已经知道 $T^* = T^*(S)$ 是无偏的，并且

$$\mathrm{var}_\theta(T^*) \le \mathrm{var}_\theta(T) \quad \forall \theta.$$

若 $T'(S)$ 是另一个 $g(\theta)$ 的无偏估计量，则有:

$$\mathbb{E}_\theta(T^*(S) - T'(S)) = 0, \quad \forall \theta.$$

由于 S 是完全统计量，这意味着:

$$T^* = T', \quad P_\theta - 几乎处处.$$

例子 5.3.1 Poisson 分布下的 UMVU 估计量

设 $X_1, \ldots, X_n$ 是独立同分布的 Poisson (θ) 随机变量.目标是估计早期失效的概率:

$$g(\theta) := e^{-\theta}.$$

一个无偏估计量为:

$$T(X_1, \ldots, X_n) := 1\{X_1 = 0\}.$$

一个充分统计量为:

$$S := \sum_{i=1}^{n} X_i.$$

检查 S 是否完全

S 的分布为 $\mathrm{Poisson}(n\theta)$，因此对于任意函数 h，

$$\mathbb{E}_\theta h(S) = \sum_{k=0}^{\infty} e^{-n\theta} \frac{(n\theta)^k}{k!} h(k).$$

若

$$\mathbb{E}_\theta h(S) = 0 \quad \forall \theta,$$

则有

$$\sum_{k=0}^{\infty} \frac{(n\theta)^k}{k!} h(k) = 0 \quad \forall \theta.$$

考虑一个在零点可展开的函数 f，其 Taylor 展开为：

$$f(x)=\sum_{k=0}^{\infty}\frac{x^k}{k!}f^{(k)}(0).$$

如果左边对于所有 x 都为 0，则 $f\equiv 0$，从而 $f^{(k)}(0)=0$ 对所有 k 都成立.令 $h(k)$ 扮演 $f^{(k)}(0)$ 的角色，而 $n\theta$ 扮演 x 的角色，可以得出 $h(k)=0$ 对所有 k 成立，即 S 是完全统计量.

计算 T^*

根据 Lehmann-Scheffé 引理，$T^*:=E(T\mid S)$ 是 UMVU.现在计算 T^*：

$$\begin{aligned}P(T=1\mid S=s)&=P(X_1=0\mid S=s)\\&=\frac{e^{-\theta}e^{-(n-1)\theta}[(n-1)\theta]^s/s!}{e^{-n\theta}(n\theta)^s/s!}\\&=\left(\frac{n-1}{n}\right)^s.\end{aligned}$$

因此，

$$T^*=\left(\frac{n-1}{n}\right)^S$$

是 UMVU.

例子 5.3.2 均匀分布的 UMVU 估计

设 $X_1,\dots,X_n$ 是独立同分布的 $[0,\theta]$ 上的均匀分布随机变量，目标是估计 $g(\theta):=\theta$.我们知道 $S:=\max\{X_1,\dots,X_n\}$ 是充分统计量（见例子 4.2.1）.S 的分布函数为：

$$F_S(s)=P_\theta\left(\max\{X_1,\dots,X_n\}\le s\right)=\left(\frac{s}{\theta}\right)^n,\quad 0\le s\le\theta.$$

因此，其概率密度函数为：

$$f_S(s)=\frac{ns^{n-1}}{\theta^n},\quad 0\le s\le\theta.$$

对于任意可测函数 h，有：

$$\mathbb{E}_\theta h(S)=\int_0^\theta h(s)\frac{ns^{n-1}}{\theta^n}\,ds.$$

若

$$\mathbb{E}_\theta h(S)=0\quad\forall\theta,$$

则必须满足

$$\int_0^\theta h(s)s^{n-1}\,ds = 0 \quad \forall\theta.$$

对 θ 求导得：

$$h(\theta)\theta^{n-1} = 0 \quad \forall\theta,$$

这表明 $h \equiv 0$.因此，S 是完全统计量.

接下来需要找到一个仅依赖于 S 且无偏的统计量.我们计算：

$$\mathbb{E}_\theta S = \int_0^\theta s\frac{ns^{n-1}}{\theta^n}\,ds = \frac{n}{n+1}\theta.$$

因此，S 本身不是无偏的，它略小于 θ.但可以通过简单调整使其无偏：取

$$T^* = \frac{n+1}{n}S.$$

根据 Lehmann-Scheffé 引理，T^* 是 UMVU.

5.4 指数族的完全性

对于指数族，当参数空间与充分统计量的维度一致时，充分统计量是完全的.这一点在以下引理中得到了形式化的陈述.我们略去证明.

引理 5.4.1

对于 $\theta \in \Theta$，

$$p_\theta(x) = \exp\left[\sum_{j=1}^k c_j(\theta)T_j(x) - d(\theta)\right]h(x).$$

考虑集合：

$$\mathcal{C} := \{(c_1(\theta), \ldots, c_k(\theta)) : \theta \in \Theta\} \subset \mathbb{R}^k.$$

若 $\mathcal{C}$ 真正是 k 维的（即，不低于 k 维，例如包含 $\mathbb{R}^k$ 中的一个开球或开立方体 $\prod_{j=1}^k (a_j, b_j)$），则 $S := (T_1, \ldots, T_k)$ 是完全的.

例子 5.4.1 Gamma 分布的完全性

设 $X_1, \ldots, X_n$ 是独立同分布的 $\Gamma(k, \lambda)$ 随机变量，其中 k 和 λ 都未知，即 $\theta := (k, \lambda)$，且 $\Theta := \mathbb{R}_+^2$.$\Gamma(k, \lambda)$ 分布的密度函数为：

$$f(z) = \frac{\lambda^k}{\Gamma(k)} e^{-\lambda z} z^{k-1}, \quad z > 0.$$

因此，

$$p_\theta(x) = \exp\left[-\lambda \sum_{i=1}^n x_i + (k-1) \sum_{i=1}^n \log x_i - d(\theta)\right] h(x),$$

其中，

$$d(k, \lambda) = -nk \log \lambda + n \log \Gamma(k),$$

且

$$h(x) = 1\{x_i > 0, i = 1, \ldots, n\}.$$

由此可知：

$$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n \log X_i \right)$$

是充分且完全的统计量.

例子 5.4.2 两个正态样本的完全性

考虑两个来自正态分布的独立样本：$X_1, \ldots, X_n$ 是独立同分布的 $\mathcal{N}(\mu, \sigma^2)$，而 $Y_1, \ldots, Y_m$ 是独立同分布的 $\mathcal{N}(\nu, \tau^2)$.

情况 i

若 $\theta = (\mu, \nu, \sigma^2, \tau^2) \in \mathbb{R}^2 \times \mathbb{R}_+^2$，容易验证以下统计量：

$$S := \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{j=1}^m Y_j, \sum_{j=1}^m Y_j^2 \right)$$

是充分且完全的.

情况 ii

若 μ, σ^2 和 τ^2 是未知的，且 $\nu = \mu$，则 S 仍然是充分的.但是可以证明，S 在此情况下不是完全统计量.

一个难题：是否存在充分且完全的统计量?

5.5 Cramér-Rao 下界

设 $\{P_\theta : \theta \in \Theta\}$ 是定义在样本空间 $\mathcal{X}$ 上的一族分布，且由一个 σ -有限测度 ν 主导.我们用以下方式定义其密度函数：

$$p_\theta := \frac{dP_\theta}{d\nu}, \quad \theta \in \Theta.$$

在本节中，假设 Θ 是一维开区间（对高维参数空间的扩展将在下一节处理）.

我们将引入以下两个条件：

条件 I
集合

$$A := \{x : p_\theta(x) > 0\}$$

对所有 θ 不依赖于 θ.

条件 II (在 L_2 中的可微性)
对于所有 θ，若存在函数 $s_\theta : \mathcal{X} \to \mathbb{R}$ 满足

$$I(\theta) := E_\theta s_\theta^2(X) < \infty,$$

则以下极限成立：

$$\lim_{h \to 0} E_\theta \left(\frac{p_{\theta+h}(X) - p_\theta(X)}{h p_\theta(X)} - s_\theta(X) \right)^2 = 0.$$

定义 5.5.1
若条件 I 和 II 满足，我们称 s_θ 为得分函数，$I(\theta)$ 为 Fisher 信息量.

通过对比定义 4.7.1 可知，在 $\theta \mapsto p_\theta$ 可微并满足“正则性条件”时，这些定义一致.回顾引理 4.7.1 中，我们假设了一些未明确说明的“正则性条件”.下面我们将给出该引理第一部分的严格证明.

引理 5.5.1
若条件 I 和 II 成立，则

$$E_\theta s_\theta(X) = 0, \quad \forall \theta.$$

证明
在 P_θ 下，我们仅需考虑 x 满足 $p_\theta(x) > 0$ 的情形，即可以安全地将 $p_\theta(x)$ 作为分母，而不必担心除以零.

注意到：

$$E_\theta \left(\frac{p_{\theta+h}(X) - p_\theta(X)}{p_\theta(X)} \right) = \int_A (p_{\theta+h} - p_\theta)\, d\nu = 0,$$

因为密度函数的积分为 1，且 $p_{\theta+h}$ 和 p_θ 在 A 外均为 0.因此，

$$
\begin{aligned}
\left|E_\theta s_\theta(X)\right|^2 &= \left|E_\theta\left(\frac{p_{\theta+h}(X)-p_\theta(X)}{hp_\theta(X)}-s_\theta(X)\right)\right|^2 \\
&\leq E_\theta\left(\frac{p_{\theta+h}(X)-p_\theta(X)}{hp_\theta(X)}-s_\theta(X)\right)^2 \to 0.
\end{aligned}
$$

第 5 章：偏差、方差与 Cramér-Rao 下界（续）

注解： 因此，$I(\theta)=\operatorname{var}_\theta\left(s_\theta(X)\right)$.

备注： 如之前所述，如果 $p_\theta(x)$ 对所有 x 可微，那么我们可以令（在正则条件下）：

$$
s_\theta(x):=\frac{d}{d\theta}\log p_\theta(x)=\frac{\dot{p}_\theta(x)}{p_\theta(x)}
$$

其中，

$$
\dot{p}_\theta(x):=\frac{d}{d\theta}p_\theta(x).
$$

备注： 假设 $X_1,\ldots,X_n$ 独立同分布于密度 p_θ，且 $s_\theta=\dot{p}_\theta/p_\theta$ 存在.则联合密度为：

$$
\mathbf{p}_\theta(\mathbf{x})=\prod_{i=1}^n p_\theta\left(x_i\right)
$$

因此（在条件 I 和 II 下），对于 n 个观测值的得分函数为：

$$
\mathbf{s}_\theta(\mathbf{x})=\sum_{i=1}^n s_\theta\left(x_i\right)
$$

n 个观测值的 Fisher 信息为：

$$
\mathbf{I}(\theta)=\operatorname{var}_\theta\left(\mathbf{s}_\theta(\mathbf{X})\right)=\sum_{i=1}^n \operatorname{var}_\theta\left(s_\theta\left(X_i\right)\right)=nI(\theta)
$$

定理 5.5.1 （Cramér-Rao 下界）

假设条件 I 和 II 满足，且 T 是 $g(\theta)$ 的无偏估计量，方差有限.那么 $g(\theta)$ 可导，且导数 $\dot{g}(\theta):=dg(\theta)/d\theta$ 满足：

$$
\dot{g}(\theta)=\operatorname{cov}\left(T,s_\theta(X)\right)
$$

此外，

$$
\operatorname{var}_\theta(T)\geq\frac{\dot{g}^2(\theta)}{I(\theta)},\forall\theta
$$

证明：

首先，展示 $g(\theta)$ 的可导性.由于 T 是无偏的，有：

$$\begin{aligned}\frac{g(\theta+h)-g(\theta)}{h} &= \frac{\mathbb{E}_{\theta+h}T(X)-\mathbb{E}_{\theta}T(X)}{h}\\ &= \frac{1}{h}\int T\left(p_{\theta+h}-p_{\theta}\right)d\nu\\ &= \mathbb{E}_{\theta}T(X)\frac{p_{\theta+h}(X)-p_{\theta}(X)}{hp_{\theta}(X)}.\end{aligned}$$

继续分解：

$$\begin{aligned}\mathbb{E}_{\theta}T(X)\frac{p_{\theta+h}(X)-p_{\theta}(X)}{hp_{\theta}(X)} &= \mathbb{E}_{\theta}T(X)\left(\frac{p_{\theta+h}(X)-p_{\theta}(X)}{hp_{\theta}(X)}-s_{\theta}(X)\right)\\ &\quad+\mathbb{E}_{\theta}T(X)s_{\theta}(X).\end{aligned}$$

根据 Cauchy-Schwarz 不等式，有：

$$\begin{aligned}&\left|\mathbb{E}_{\theta}\left(T(X)-g_{\theta}\right)\left(\frac{p_{\theta+h}(X)-p_{\theta}(X)}{hp_{\theta}(X)}-s_{\theta}(X)\right)\right|^2\\ &\leq \operatorname{var}_{\theta}(T)\mathbb{E}_{\theta}\left(\frac{p_{\theta+h}(X)-p_{\theta}(X)}{hp_{\theta}(X)}-s_{\theta}(X)\right)^2\to 0.\end{aligned}$$

因此，

$$\dot{g}(\theta)=\mathbb{E}_{\theta}T(X)s_{\theta}(X)=\operatorname{cov}_{\theta}\left(T,s_{\theta}(X)\right).$$

最后，不等式由 Cauchy-Schwarz 推得：

$$\begin{aligned}\dot{g}^2(\theta) &= \left(\operatorname{cov}_{\theta}\left(T,s_{\theta}(X)\right)\right)^2\\ &\leq \operatorname{var}_{\theta}(T)\operatorname{var}_{\theta}\left(s_{\theta}(X)\right)=\operatorname{var}_{\theta}(T)I(\theta).\end{aligned}$$

定义 5.5.2： 我们称 $\dot{g}^2(\theta)/I(\theta),\theta\in\Theta$ 为 $g(\theta)$ 的 Cramér-Rao 下界（CRLB）.

例子 5.5.1：指数分布的 CRLB

设 $X_1,\ldots,X_n$ 独立同分布于指数分布 $\text{Exponential}(\theta),\theta>0$.单个观测值的密度为：

$$p_{\theta}(x)=\theta e^{-\theta x},\quad x>0.$$

令 $g(\theta):=1/\theta$，$T:=\bar{X}$.可以验证 T 是无偏的，且方差为 $\operatorname{var}_{\theta}(T)=1/(n\theta^2)$.

计算 CRLB.令 $g(\theta)=1/\theta$，有 $\dot{g}(\theta)=-1/\theta^2$.此外，

$$\log p_{\theta}(x)=\log\theta-\theta x,$$

所以

$$s_\theta(x) = 1/\theta - x,$$

因此

$$I(\theta) = \mathrm{var}_\theta(X) = \frac{1}{\theta^2}.$$

n 个观测值的 CRLB 为：

$$\frac{\dot{g}^2(\theta)}{nI(\theta)} = \frac{1}{n\theta^2}.$$

换句话说，T 达到了 CRLB.

例子 5.5.2：泊松分布的 CRLB

假设 $X_1, \ldots, X_n$ 独立同分布于泊松分布 $\mathrm{Poisson}(\theta), \theta > 0$.则

$$\log p_\theta(x) = -\theta + x \log \theta - \log x!.$$

对于泊松分布的得分函数 $s_\theta(x)$，我们有：

$$s_\theta(x) = \frac{\partial}{\partial \theta} \log p_\theta(x) = -1 + \frac{x}{\theta}$$

Fisher 信息为：

$$I(\theta) = \mathbb{E}_\theta[s_\theta(X)^2] = \mathbb{E}_\theta\left[\left(-1 + \frac{X}{\theta}\right)^2\right].$$

展开并计算期望值：

$$I(\theta) = \mathbb{E}_\theta\left[1 - \frac{2X}{\theta} + \frac{X^2}{\theta^2}\right].$$

利用泊松分布的性质 $\mathbb{E}_\theta(X) = \theta$ 和 $\mathrm{var}_\theta(X) = \theta$，得到：

$$I(\theta) = 1 - \frac{2\theta}{\theta} + \frac{\theta + \theta^2}{\theta^2} = \frac{1}{\theta}.$$

因此，n 个观测值的 Fisher 信息为：

$$\mathbf{I}(\theta) = nI(\theta) = \frac{n}{\theta}.$$

设 $g(\theta) = \theta$，则 $\dot{g}(\theta) = 1$.Cramér-Rao 下界为：

$$\frac{\dot{g}^2(\theta)}{\mathbf{I}(\theta)} = \frac{1}{n/\theta} = \frac{\theta}{n}.$$

这说明，对于泊松分布参数 θ 的无偏估计量，其方差的理论下界为 θ/n.泊松分布样本均值 $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ 是无偏估计量，且其方差恰好等于 θ/n，因此达到了 Cramér-Rao 下界.

这进一步说明了 CRLB 在评估估计量效率方面的重要性，例如样本均值在这些分布下的最优性.

第 5 章：偏差、方差与 Cramér-Rao 下界（续）

因此，有

$$s_\theta(x) = -1 + \frac{x}{\theta}$$

由此，

$$I(\theta) = \mathrm{var}_\theta\left(\frac{X}{\theta}\right) = \frac{\mathrm{var}_\theta(X)}{\theta^2} = \frac{1}{\theta}$$

容易验证，样本均值 $\bar{X}$ 达到了估计 θ 的 Cramér-Rao 下界.

现在设 $g(\theta) := \mathrm{e}^{-\theta}$，则其 UMVU 估计量为：

$$T := \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^n X_i}$$

为了计算其方差，首先计算 $\mathbb{E}_\theta T^2$：

$$\begin{aligned}
\mathbb{E}_\theta T^2 &= \sum_{k=0}^{\infty} \left(1 - \frac{1}{n}\right)^{2k} \frac{(n\theta)^k}{k!} \mathrm{e}^{-n\theta} \\
&= \mathrm{e}^{-n\theta} \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{(n-1)^2\theta}{n}\right)^k \\
&= \mathrm{e}^{-n\theta} \exp\left[\frac{(n-1)^2\theta}{n}\right] \\
&= \exp\left[\frac{(1-2n)\theta}{n}\right]
\end{aligned}$$

因此，

$$\begin{aligned}
\mathrm{var}_\theta(T) &= \mathbb{E}_\theta T^2 - (\mathbb{E}_\theta T)^2 = \mathbb{E}_\theta T^2 - \mathrm{e}^{-2\theta} \\
&= \mathrm{e}^{-2\theta}\left(\mathrm{e}^{\theta/n} - 1\right) \\
&\begin{cases} > \frac{\theta \mathrm{e}^{-2\theta}}{n} \\ \approx \frac{\theta \mathrm{e}^{-2\theta}}{n}, \text{当 } n \text{ 足够大时.} \end{cases}
\end{aligned}$$

由于 $\dot{g}(\theta) = -\mathrm{e}^{-\theta}$，Cramér-Rao 下界为：

$$\frac{\dot{g}^2(\theta)}{nI(\theta)} = \frac{\theta \mathrm{e}^{-2\theta}}{n}$$

由此可知，T 未达到 Cramér-Rao 下界，但当 n 足够大时，差距很小.

5.6 Cramér-Rao 下界与指数族

以下结果表明，Cramér-Rao 下界仅能在指数族内达到，因此在有限的上下文中才紧.

引理 5.6.1

假设条件 I 和 II 满足，并且 $s_\theta = \dot{p}_\theta / p_\theta$.如果 T 是 $g(\theta)$ 的无偏估计量，并且 T 达到了 Cramér-Rao 下界，则 $\{P_\theta : \theta \in \Theta\}$ 构成一维指数族.即，存在函数 $c(\theta), d(\theta)$ 和 $h(x)$，使得：

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x), \quad x \in \mathcal{X}$$

此外，$c(\theta)$ 和 $d(\theta)$ 可导，其导数分别为 $\dot{c}(\theta)$ 和 $\dot{d}(\theta)$.并且有如下关系：

$$g(\theta) = \dot{d}(\theta)/\dot{c}(\theta), \quad \forall \theta$$

证明

根据定理 5.5，如果 T 达到了 Cramér-Rao 下界，则有：

$$\mathrm{var}_\theta(T) = \frac{|\mathrm{cov}_\theta\left(T, s_\theta(X)\right)|^2}{\mathrm{var}_\theta\left(s_\theta(X)\right)}$$

即，T 和 $s_\theta(X)$ 的相关系数为 ± 1.因此，存在关于 θ 的常数 $a(\theta)$ 和 $b(\theta)$，使得：

$$s_\theta(X) = a(\theta)\, T(X) - b(\theta) \tag{5.1}$$

因为 $s_\theta = \frac{\dot{p}_\theta}{p_\theta} = \frac{d \log p_\theta}{d\theta}$，可以进行积分，得到：

$$\log p_\theta(x) = c(\theta)T(x) - d(\theta) + \tilde{h}(x)$$

其中 $\dot{c}(\theta) = a(\theta), \dot{d}(\theta) = b(\theta)$，且 $\tilde{h}(x)$ 与 θ 无关.因此，

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$$

其中 $h(x) = \exp[\tilde{h}(x)]$.

此外，由等式 (5.1) 可知：

$$\mathbb{E}_\theta s_\theta(X) = a(\theta)\mathbb{E}_\theta T - b(\theta) = a(\theta)g(\theta) - b(\theta)$$

因为 $\mathbb{E}_\theta s_\theta(X) = 0$，所以有：

$$g(\theta) = b(\theta)/a(\theta)$$

这完成了证明.

第 5.7 节：高维扩展

随机向量的期望和协方差矩阵

令 $Z \in \mathbb{R}^k$ 为 k 维随机向量，其期望为 EZ，是一个 k 维向量，协方差矩阵定义为

$$\Sigma := \mathrm{Cov}(Z) := E[ZZ'] - (EZ)(EZ'),$$

这是一个 $k \times k$ 的矩阵，其中对角线包含方差，非对角线包含协方差.需要注意的是，Σ 是半正定的：对于任意向量 $a \in \mathbb{R}^k$，有

$$\mathrm{var}(a'Z) = a'\Sigma a \geq 0.$$

矩阵代数简介

令 V 为对称矩阵，如果 V 是正定（或半正定）的，记为 $V > 0$（或 $V \geq 0$）.此时，有 $V = W^2$，其中 W 也是正定（或半正定）的.

辅助引理 假设 $V > 0$，则有

$$\max_{a \in \mathbb{R}^p} \frac{|a'c|^2}{a'Va} = c'V^{-1}c.$$

证明 令 $V = W^2$，并令 $b := Wa, d := W^{-1}c$，则 $a'Va = b'b = \|b\|^2$ 且 $a'c = b'd$.根据 Cauchy-Schwarz 不等式，

$$\max_{b \in \mathbb{R}^p} \frac{|b'd|^2}{\|b\|^2} = \|d\|^2 = d'd = c'V^{-1}c.$$

高维 Cramér-Rao 下界

我们现在陈述高维情形下的 Cramér-Rao 下界，为简化叙述，假设所有必要的导数存在，并且在适当的位置可以交换微分与积分.

设参数空间 $\Theta \subset \mathbb{R}^k$，给定函数

$$g : \Theta \to \mathbb{R}.$$

定义偏导数的向量为

$$\dot{g}(\theta) := \begin{pmatrix} \frac{\partial g(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial g(\theta)}{\partial \theta_k} \end{pmatrix}.$$

记得分向量为

$$s_\theta(\cdot) := \begin{pmatrix} \frac{\partial \log p_\theta}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log p_\theta}{\partial \theta_k} \end{pmatrix},$$

Fisher 信息矩阵为

$$I(\theta) = E_\theta[s_\theta(X)s'_\theta(X)] = \mathrm{Cov}_\theta(s_\theta(X)).$$

定理 5.7.1

令 T 为 $g(\theta)$ 的无偏估计量，则在正则性条件下，

$$\mathrm{var}_\theta(T) \geq \dot{g}(\theta)' I(\theta)^{-1} \dot{g}(\theta).$$

证明 与一维情形类似，可证明对 $j = 1, \ldots, k$，

$$\dot{g}_j(\theta) = \mathrm{cov}_\theta\left(T, s_{\theta,j}(X)\right).$$

因此，对于任意 $a \in \mathbb{R}^k$，

$$\begin{aligned} |a'\dot{g}(\theta)|^2 &= |\,\mathrm{cov}_\theta(T, a's_\theta(X))|^2 \\ &\leq \mathrm{var}_\theta(T)\,\mathrm{var}_\theta(a's_\theta(X)) \\ &= \mathrm{var}_\theta(T) a' I(\theta) a. \end{aligned}$$

结合辅助引理，有

$$\mathrm{var}_\theta(T) \geq \max_{a \in \mathbb{R}^k} \frac{|a'\dot{g}(\theta)|^2}{a'I(\theta)a} = \dot{g}'(\theta) I(\theta)^{-1} \dot{g}(\theta).$$

推论 5.7.1

高维参数感兴趣时，无偏估计量也可得到下界.例如，令 $g(\theta) := \theta = (\theta_1, \ldots, \theta_k)'$，假设 $T \in \mathbb{R}^k$ 是 θ 的无偏估计量：$E_\theta T = \theta, \forall \theta$.则对于任意 $a \in \mathbb{R}^k$，$a'T$ 是 $a'\theta$ 的无偏估计量.由于 $a'\theta$ 的导数为 a，CRLB 给出

$$\mathrm{var}_\theta(a'T) \geq a' I(\theta)^{-1} a.$$

但

$$\operatorname{var}_\theta(a'T) = a' \operatorname{Cov}_\theta(T) a,$$

因此，对于所有 a，

$$a' \operatorname{Cov}_\theta(T) a \geq a' I(\theta)^{-1} a,$$

即 $\operatorname{Cov}_\theta(T) \geq I(\theta)^{-1}$，换言之，$\operatorname{Cov}_\theta(T) - I(\theta)^{-1}$ 是半正定的.

第 6 章

检验与置信区间

6.1 插曲：分位数函数

设 F 为定义在 $\mathbb{R}$ 上的分布函数，则 F 是右连续且左极限存在的（cadlag: continue à droite, limite à gauche）.定义分位数函数如下：

$$q^F_{\sup}(u) := \sup\{x : F(x) \leq u\},$$

以及

$$q^F_{\inf}(u) := \inf\{x : F(x) \geq u\} := F^{-1}(u).$$

以下性质成立：

$$F\left(q^F_{\inf}(u)\right) \geq u,$$

且对于任意 $h > 0$，

$$F\left(q^F_{\sup}(u) - h\right) \leq u.$$

因此，

$$F\left(q^F_{\sup}(u)-\right) := \lim_{h \downarrow 0} F\left(q^F_{\sup}(u) - h\right) \leq u.$$

6.2 如何构造检验

考虑一个模型类

$$\mathcal{P} := \{P_\theta : \theta \in \Theta\}.$$

进一步，定义一个空间 Γ 和一个映射

$$g : \Theta \to \Gamma, \quad g(\theta) := \gamma,$$

其中 γ 是感兴趣的参数.

定义 6.2.1

令 $\gamma_0 \in \Gamma$ 和 $\alpha \in [0,1]$.对于假设

$$H_0 : \gamma = \gamma_0,$$

水平为 α 的（非随机化）检验是一个统计量 $\phi(X, \gamma_0) \in \{0,1\}$，满足：

$$P_\theta(\phi(X,\gamma_0) = 1) \le \alpha, \quad \forall \theta \in \{\vartheta : g(\vartheta) = \gamma_0\}.$$

通常我们会省略 ϕ 对 γ_0 的依赖性，记为 $\phi(X) := \phi(X, \gamma_0)$.

注意

通常检验 ϕ 基于一个检验统计量 T，即形如：

$$\phi(X) = \begin{cases} 1 & \text{若 } T(X) > c \\ 0 & \text{否则} \end{cases}$$

其中 c 称为**临界值**.

检验的构造

为了检验 $H_{\gamma_0} : \gamma = \gamma_0$，我们寻找一个枢轴量（pivot），这是一个关于数据 $\mathbf{X}$ 和参数 γ 的函数 $Z(\mathbf{X}, \gamma)$，满足对任意 $\theta \in \Theta$，

$$\mathbb{P}_\theta(Z(\mathbf{X}, g(\theta)) \le \cdot) =: G(\cdot),$$

其分布 G 不依赖于 θ.

注意 找到枢轴量并非总是可能.然而，如果存在枢轴量 $Z(\mathbf{X}, \gamma)$ 且其分布为 G，我们可以计算其分位数函数：

$$q_L := q^G_{\sup}\left(\frac{\alpha}{2}\right), \quad q_R := q^G_{\inf}\left(1 - \frac{\alpha}{2}\right).$$

定义检验为：

$$\phi(\mathbf{X}, \gamma_0) := \begin{cases} 1 & \text{若 } Z(\mathbf{X}, \gamma_0) \notin [q_L, q_R] \\ 0 & \text{否则} \end{cases}$$

此时检验对 H_{γ_0} 的水平为 α，其中 $\gamma_0 = g(\theta_0)$：

$$\begin{aligned} &\mathbb{P}_{\theta_0}\left(\phi(\mathbf{X}, g(\theta_0)) = 1\right) = P_{\theta_0}\left(Z(\mathbf{X}, g(\theta_0)) > q_R\right) + P_{\theta_0}\left(Z(\mathbf{X}, g(\theta_0)) < q_L\right) \\ &= 1 - G(q_R) + G(q_L-) \le 1 - \left(1 - \frac{\alpha}{2}\right) + \frac{\alpha}{2} = \alpha. \end{aligned}$$

渐近枢轴量

令 $Z_n(X_1,\dots,X_n,\gamma)$ 为定义在每个样本量 n 上的一个关于数据和参数的函数.如果对任意 $\theta\in\Theta$，

$$\lim_{n\to\infty}\mathbb{P}_\theta(Z_n(X_1,\dots,X_n,\gamma)\le\cdot)=G(\cdot),$$

其中 limit G 不依赖于 θ，则称 $Z_n(X_1,\dots,X_n,\gamma)$ 为渐近枢轴量（asymptotic pivot）.

例子 6.2.1 位移模型

作为例子，重新考虑位移模型（参见 1.3 节）.设

$$\Theta:=\{\theta=(\mu,F_0):\mu\in\mathbb{R},F_0\in\mathcal{F}_0\},$$

其中 $\mathcal{F}_0$ 是对称分布集合的一个子集（参见 (1.2)）.令 $\hat{\mu}$ 为一个等变估计量（equivariant estimator），即 $\hat{\mu}-\mu$ 的分布与 μ 无关（关于等变性的形式定义参见第 9 章）.

- **如果** $\mathcal{F}_0:=\{F_0\}$ 是一个单一分布（即 F_0 已知），我们取枢轴量 $Z(\mathbf{X},\mu):=\hat{\mu}-\mu$.由于等变性，此枢轴量的分布 G 仅依赖于 F_0.
- **如果** $\mathcal{F}_0:=\{F_0(\cdot)=\Phi(\cdot/\sigma):\sigma>0\}$，我们选择 $\hat{\mu}:=\bar{X}_n$，其中 $\bar{X}_n=\sum_{i=1}^n X_i/n$ 为样本均值.作为枢轴量，我们取：

  $$Z(\mathbf{X},\mu):=\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n},$$

  其中 $S_n^2=\sum_{i=1}^n(X_i-\bar{X})^2/(n-1)$ 是样本方差.此时 G 是自由度为 $n-1$ 的学生 t 分布.
- **如果** $\mathcal{F}_0:=\{F_0\}$ 是对称且在 $x=0$ 连续的分布，我们令枢轴量为符号检验统计量：

  $$Z(\mathbf{X},\mu):=\sum_{i=1}^n 1\{X_i\ge\mu\}.$$

  此时 G 是参数为 $p=1/2$ 的 $\mathrm{Binomial}(n,p)$ 分布.
- **进一步假设** $X_1,\dots,X_n$ 是无限序列的前 n 个独立同分布随机变量，并且：

  $$\mathcal{F}_0:=\{F_0:\int xdF_0(x)=0,\quad\int x^2dF_0(x)<\infty\}.$$

  那么：

  $$Z_n(X_1,\dots,X_n,\mu):=\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n}$$

  是一个渐近枢轴量，其极限分布为 $G=\Phi$.

6.3 置信集与检验的等价性

定义 6.3.1

对于 γ，基于数据 $\mathbf{X}=(X_1,\ldots,X_n)$ 的子集 $I=I(\mathbf{X})\subset\Gamma$ 称为 γ 在 $1-\alpha$ 水平下的置信集，如果：

$$\mathbb{P}_\theta(\gamma\in I)\geq 1-\alpha,\quad \forall\theta\in\Theta.$$

置信区间的形式为：

$$I:=[\underline{\gamma},\bar{\gamma}],$$

其中边界 $\underline{\gamma}=\underline{\gamma}(\mathbf{X})$ 和 $\bar{\gamma}=\bar{\gamma}(\mathbf{X})$ 仅依赖于数据 $\mathbf{X}$.

对于每个 $\gamma_0\in\mathbb{R}$，令 $\phi(\mathbf{X},\gamma_0)\in\{0,1\}$ 为检验假设 $H_{\gamma_0}:\gamma=\gamma_0$ 的一个 α 水平的检验.即当且仅当 $\phi(\mathbf{X},\gamma_0)=1$ 时拒绝 H_{γ_0}，并满足：

$$\mathbb{P}_{\theta:\gamma=\gamma_0}(\phi(\mathbf{X},\gamma_0)=1)\leq\alpha.$$

那么：

$$I(\mathbf{X}):=\{\gamma:\phi(\mathbf{X},\gamma)=0\}$$

是 γ 的一个 $1-\alpha$ 置信集.

反过来，如果 $I(\mathbf{X})$ 是 γ 的一个 $1-\alpha$ 置信集，则对于任意 γ_0，定义的检验：

$$\phi(\mathbf{X},\gamma_0)=\begin{cases}1 & \text{如果 } \gamma_0\notin I(\mathbf{X})\\ 0 & \text{否则}\end{cases}$$

是 H_{γ_0} 的 α 水平检验.

6.4 置信区间与检验的比较

比较置信区间时，目标通常是选取在平均长度上最短的区间（保持水平为 $1-\alpha$）.对于检验，目标是选取具有最大功效的检验.回忆：在 θ 处（$g(\theta)\neq\gamma_0$）检验 $\phi(\mathbf{X},\gamma_0)$ 的功效为：

$$P_\theta(\phi(\mathbf{X},\gamma_0)=1).$$

6.5 例子：双样本问题

考虑关于体重增减的以下数据.对照组 x 保持常规饮食，而处理组 y 采用了一种防止体重增加的特殊饮食.研究的目的是检验该饮食是否有效.

control group x	treatment group y	rank(x)	rank(y)
5	6	7	8
0	-5	3	2
16	-6	10	1
2	1	5	4
9	4	9	6
—— +	—— +		
32	0		

表格 2

令 $n(m)$ 为对照组 x（处理组 y）的样本量，组 $x(y)$ 的均值记为 $\bar{x}(\bar{y})$.平方和为：

$$SS_x := \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad SS_y := \sum_{j=1}^{m}(y_j - \bar{y})^2.$$

在本研究中，$n = m = 5$，且 $\bar{x} = 6.4, \bar{y} = 0, SS_x = 161.2, SS_y = 114$.秩 $\mathrm{rank}(x)$ 和 $\mathrm{rank}(y)$ 是将 $n + m$ 数据合并排序后的秩值（例如 $y_3 = -6$ 是最小的观测值，因此 $\mathrm{rank}(y_3) = 1$）.

假设数据为两个独立样本的观测值，即 $\mathbf{X} = (X_1, \ldots, X_n)$ 和 $\mathbf{Y} = (Y_1, \ldots, Y_m)$，其中 $X_1, \ldots, X_n$ 独立同分布，分布函数为 F_X；$Y_1, \ldots, Y_m$ 独立同分布，分布函数为 F_Y.分布函数 F_X 和 F_Y 可以完全或部分未知.

检验问题为：

$$H_0 : F_X = F_Y,$$

与单侧或双侧备择假设的比较.

6.5.1 Student 检验

经典的双样本 Student 检验基于假设数据来自正态分布.此外，假设 F_X 和 F_Y 的方差相等.因此，

$$(F_X, F_Y) \in \left\{F_X = \Phi\left(\frac{\cdot - \mu}{\sigma}\right), F_Y = \Phi\left(\frac{\cdot - (\mu + \gamma)}{\sigma}\right) : \mu \in \mathbb{R}, \sigma > 0, \gamma \in \Gamma\right\}.$$

其中，$\Gamma \supset \{0\}$ 是考虑的均值偏移范围，例如 $\Gamma = \mathbb{R}$（双侧情形），以及 $\Gamma = (-\infty, 0]$（单侧情形）.检验问题简化为：

$$H_0 : \gamma = 0.$$

我们现在寻找一个枢轴量 $Z(\mathbf{X}, \mathbf{Y}, \gamma)$.定义样本均值为：

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} := \frac{1}{m}\sum_{j=1}^{m} Y_j,$$

以及合并样本方差为：

$$S^2 := \frac{1}{m+n-2}\left\{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + \sum_{j=1}^{m}\left(Y_j - \bar{Y}\right)^2\right\}.$$

注意，$\bar{X}$ 的期望为 μ，方差为 σ^2/n；$\bar{Y}$ 的期望为 $\mu+\gamma$，方差为 σ^2/m.因此，$\bar{Y} - \bar{X}$ 的期望为 γ，方差为：

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\left(\frac{n+m}{nm}\right).$$

正态分布假设意味着：

$$\bar{Y} - \bar{X} \text{ 服从 } \mathcal{N}\left(\gamma, \sigma^2\left(\frac{n+m}{nm}\right)\right).$$

因此，

$$\sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y} - \bar{X} - \gamma}{\sigma}\right) \text{ 服从 } \mathcal{N}(0,1).$$

为了构造枢轴量，我们用估计量 S 替代未知的 σ：

$$Z(\mathbf{X}, \mathbf{Y}, \gamma) := \sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y} - \bar{X} - \gamma}{S}\right).$$

实际上，$Z(\mathbf{X}, \mathbf{Y}, \gamma)$ 的分布 G 不依赖于未知参数.分布 G 是 $\text{Student}(n+m-2)$（自由度为 $n+m-2$ 的学生分布）.因此，针对 $H_0 : \gamma = 0$，检验统计量为：

$$T = T^{\text{Student}} := Z(\mathbf{X}, \mathbf{Y}, 0).$$

在单侧检验中，对于 $H_0 : \gamma = 0$ 对比 $H_1 : \gamma < 0$，水平为 α 的检验为：

$$\phi(\mathbf{X}, \mathbf{Y}) := \begin{cases} 1 & \text{若 } T < -t_{n+m-2}(1-\alpha) \\ 0 & \text{若 } T \geq -t_{n+m-2}(1-\alpha) \end{cases}$$

其中，对于 $\nu > 0$，$t_\nu(1-\alpha) = -t_\nu(\alpha)$ 是 $\text{Student}(\nu)$ 分布的 $(1-\alpha)$ 分位数.

应用于表格 2 中的数据：

取 $\alpha = 0.05$.观测值为 $\bar{x} = 6.4, \bar{y} = 0, s^2 = 34.4$.检验统计量的值为 -1.725，大于 5% 分位数 $t_8(0.05) = -1.9$.因此，我们不能拒绝 H_0.

观测值 T 的 p 值为：

$$p\text{-value} := \mathbb{P}_{\gamma=0}(T < -1.725) = 0.06.$$

因此，在本例中，p 值仅比水平 $\alpha = 0.05$ 稍大.

6.5.2 Wilcoxon 检验

本小节中，我们假设 F_X 和 F_Y 是连续分布，但其他信息未知.因此，F_X 和 F_Y 的模型类为：

$$\mathcal{F} := \{\text{所有连续分布}\}.$$

连续性假设确保所有观测值是不同的，即没有重复值.因此，我们可以将观测值按严格递增顺序排列.令 $N = n + m$，$Z_1, \ldots, Z_N$ 为合并样本：

$$Z_i := X_i, i = 1, \ldots, n; \quad Z_{n+j} := Y_j, j = 1, \ldots, m.$$

定义秩次为：

$$R_i := \text{rank}(Z_i), \quad i = 1, \ldots, N.$$

并令：

$$Z_{(1)} < \cdots < Z_{(N)}$$

表示合并样本的次序统计量（即，$Z_i = Z_{(R_i)}$ for $i = 1, \ldots, N$）.Wilcoxon 检验统计量定义为：

$$T = T^{\text{Wilcoxon}} := \sum_{i=1}^{n} R_i.$$

可以验证，该统计量 T 也可以表示为：

$$T = \#\{Y_j < X_i\} + \frac{n(n+1)}{2}.$$

例如，对于表 2 中的数据，观测值为 $T = 34$，其中：

$$\#\{y_j < x_i\} = 19, \quad \frac{n(n+1)}{2} = 15.$$

较大的 T 值表示 X_i 通常大于 Y_j，因此提供了反对 H_0 的证据.

要检查观测到的检验统计量是否与零假设兼容，我们需要了解其在 H_0 下的零分布，即在 H_0 下的分布形式.在 $H_0 : F_X = F_Y$ 下，秩次向量 $(R_1, \ldots, R_n)$ 的分布等价于从 $\{1, \ldots, N\}$ 中随机抽取 n 个不放回的数.因此，若令：

$$\mathbf{r} := (r_1, \ldots, r_n, r_{n+1}, \ldots, r_N)$$

表示 $\{1,\ldots,N\}$ 的一个排列，则有：

$$\mathbb{P}_{H_0}\left((R_1,\ldots,R_n,R_{n+1},\ldots,R_N)=\mathbf{r}\right)=\frac{1}{N!}.$$

（见定理 6.5.1）.因此，

$$\mathbb{P}_{H_0}(T=t)=\frac{\#\{\mathbf{r}:\sum_{i=1}^{n}r_i=t\}}{N!}.$$

同样，可以表示为：

$$\mathbb{P}_{H_0}(T=t)=\frac{1}{\binom{N}{n}}\#\{r_1<\cdots<r_n<r_{n+1}<\cdots<r_N:\sum_{i=1}^{n}r_i=t\}.$$

显然，T 的零分布不依赖于 F_X 或 F_Y，但依赖于样本量 n 和 m.对于小到中等大小的 n 和 m，可以查表；而对于较大的 n 和 m，可使用正态分布近似.

定理 6.5.1

该定理形式化推导了检验统计量的零分布，并证明了次序统计量和秩次是独立的.后者将在例子 4.1.4 中有意义.

对于两个随机变量 X 和 Y，当 X 和 Y 具有相同分布时，记作：

$$X\overset{\mathcal{D}}{=}Y.$$

定理 6.5.1 令 $Z_1,\ldots,Z_N$ 为服从 $\mathbb{R}$ 中连续分布 F 的独立同分布随机变量.则 $\left(Z_{(1)},\ldots,Z_{(N)}\right)$ 和 $\mathbf{R}:=(R_1,\ldots,R_N)$ 是独立的，对于所有排列 $\mathbf{r}:=(r_1,\ldots,r_N)$，

$$\mathbb{P}(\mathbf{R}=\mathbf{r})=\frac{1}{N!}.$$

证明：

令 $Z_{Q_i}:=Z_{(i)}$，$\mathbf{Q}:=(Q_1,\ldots,Q_N)$.则

$$\mathbf{R}=\mathbf{r}\Leftrightarrow\mathbf{Q}=\mathbf{r}^{-1}:=\mathbf{q},$$

其中 $\mathbf{r}^{-1}$ 是 $\mathbf{r}$ 的逆置换.对于所有排列 $\mathbf{q}$ 和所有可测映射 f，

$$f\left(Z_1,\ldots,Z_N\right)\overset{\mathcal{D}}{=}f\left(Z_{q_1},\ldots,Z_{q_N}\right).$$

因此，对于所有可测集合 $A\subset\mathbb{R}^N$ 和所有排列 $\mathbf{q}$，

$$\begin{aligned}&\mathbb{P}\left((Z_1,\ldots,Z_N)\in A,Z_1<\cdots<Z_N\right)\\=&\mathbb{P}\left((Z_{q_1},\ldots,Z_{q_N})\in A,Z_{q_1}<\cdots<Z_{q_N}\right).\end{aligned}$$

通过 $N!$ 种排列，可以得到：

$$\mathbb{P}\left((Z_{(1)},\ldots,Z_{(N)})\in A\right)=N!\mathbb{P}\left((Z_{q_1},\ldots,Z_{q_N})\in A,Z_{q_1}<\cdots<Z_{q_N}\right).$$

从而证明了次序统计量 $(Z_{(1)},\ldots,Z_{(N)})$ 和秩次 $\mathbf{R}$ 是独立的.
[1]

6.5.3 对比 Student 检验和 Wilcoxon 检验

由于 Wilcoxon 检验仅依赖于秩次，而不依赖正态分布假设，因此在数据实际上服从正态分布的情况下，Wilcoxon 检验的效能（power）通常比 Student 检验稍低.然而，这种效能的损失很小.让我们通过两种检验的相对效率来更精确地描述这一点.

设显著性水平 α 固定，效能 β 固定.令样本大小 n 和 m 相等，总样本量为 $N=2n$.假设 N^{Student} 和 N^{Wilcoxon} 分别是为了达到效能 β 而需要的样本量.考虑位置偏移备择假设，即 $F_Y(\cdot)=F_X(\cdot-\gamma)$（在我们的例子中，$\gamma<0$）.可以证明，当模型假设正态分布正确时，$N^{\text{Student}}/N^{\text{Wilcoxon}}\approx 0.95$.对于一大类分布，这一比值介于 0.85 到 ∞ 之间.

也就是说，使用 Wilcoxon 检验通常只会导致非常有限的效率损失，与 Student 检验相比甚至可能有显著的效率增益.

第 7 章

Neyman-Pearson 引理与 UMP 检验

设 $\mathcal{P}:=\{P_\theta:\theta\in\Theta\}$ 是一族概率测度.令 $\Theta_0\subset\Theta$ 和 $\Theta_1\subset\Theta$，且 $\Theta_0\cap\Theta_1=\emptyset$.基于观测值 $X\in\mathcal{X}$，分布为 $P\in\mathcal{P}$，我们考虑以下检验问题：
$H_0:\theta\in\Theta_0$
对比
$H_1:\theta\in\Theta_1$.

一个（可能随机化的）检验是一个函数 $\phi:\mathcal{X}\to[0,1]$.设 $\alpha\in[0,1]$.若满足：

$$\sup_{\theta\in\Theta_0}\mathbb{E}_\theta\phi(X)\le\alpha,$$

则称 ϕ 是水平为 α 的检验.

定义 7.0.1 若 ϕ 满足以下条件，则称其为一致最强检验（UMP，Uniformly Most Powerful）：

- ϕ 是水平为 α 的检验；
- 对于所有水平为 α 的检验 ϕ'，均有 $\mathbb{E}_\theta\phi'(X)\le\mathbb{E}_\theta\phi(X)$，对于所有 $\theta\in\Theta_1$.

7.1 Neyman-Pearson 引理

我们考虑如下检验问题：
$H_0 : \theta = \theta_0$
对比备择假设
$H_1 : \theta = \theta_1.$

定义检验 ϕ 的风险 $R(\theta, \phi)$ 为第一类和第二类错误的概率：

$$R(\theta, \phi) := \begin{cases} \mathbb{E}_\theta \phi(X), & \theta = \theta_0, \\ 1 - \mathbb{E}_\theta \phi(X), & \theta = \theta_1. \end{cases}$$

令 p_0 和 p_1 分别为 P_{θ_0} 和 P_{θ_1} 关于某个支配测度 ν 的密度（例如 $\nu = P_{\theta_0} + P_{\theta_1}$）.一个 Neyman-Pearson 检验的形式为：

$$\phi_{\mathrm{NP}} := \begin{cases} 1, & \text{若 } p_1/p_0 > c, \\ q, & \text{若 } p_1/p_0 = c, \\ 0, & \text{若 } p_1/p_0 < c, \end{cases}$$

其中 $0 \le q \le 1$，且 $0 \le c < \infty$ 是给定常数.

引理 7.1.1 Neyman-Pearson 引理

设 ϕ 是某个检验，则有：

$$R(\theta_1, \phi_{\mathrm{NP}}) - R(\theta_1, \phi) \le c\,[R(\theta_0, \phi) - R(\theta_0, \phi_{\mathrm{NP}})].$$

证明：

$$\begin{aligned} R(\theta_1, \phi_{\mathrm{NP}}) - R(\theta_1, \phi) &= \int (\phi - \phi_{\mathrm{NP}}) p_1 \\ &= \int_{p_1/p_0 > c} (\phi - \phi_{\mathrm{NP}}) p_1 + \int_{p_1/p_0 = c} (\phi - \phi_{\mathrm{NP}}) p_1 + \int_{p_1/p_0 < c} (\phi - \phi_{\mathrm{NP}}) p_1 \\ &\le c \int_{p_1/p_0 > c} (\phi - \phi_{\mathrm{NP}}) p_0 + c \int_{p_1/p_0 = c} (\phi - \phi_{\mathrm{NP}}) p_0 + c \int_{p_1/p_0 < c} (\phi - \phi_{\mathrm{NP}}) p_0 \\ &= c\,[R(\theta_0, \phi) - R(\theta_0, \phi_{\mathrm{NP}})]. \end{aligned}$$

7.2 一致最强检验

7.2.1 例子

设 $X_1, \ldots, X_n$ 是 n 个 i.i.d. 的伯努利随机变量，其中 $X \in \{0, 1\}$，成功概率为 $\theta \in (0, 1)$：

$$P_\theta(X = 1) = 1 - P_\theta(X = 0) = \theta.$$

我们考虑以下三个检验问题.在所有三种情况下，检验水平均为 $\alpha = 0.05$.

问题 1

我们希望以显著性水平 α 检验如下假设：

$$H_0 : \theta = \frac{1}{2} =: \theta_0,$$

对比备择假设：

$$H_1 : \theta = \frac{1}{4} =: \theta_1.$$

设 $T := \sum_{i=1}^{n} X_i$ 为成功次数的总和（T 是一个充分统计量），我们考虑以下随机化检验：

$$\phi(T) := \begin{cases} 1, & \text{若 } T < t_0, \\ q, & \text{若 } T = t_0, \\ 0, & \text{若 } T > t_0, \end{cases}$$

其中 $q \in (0,1)$，t_0 为检验的临界值.常数 q 和 $t_0 \in \{0, \ldots, n\}$ 被选择为满足以下条件，即当 H_0 为真时，拒绝 H_0 的概率等于 α：

$$P_{\theta_0}\left(H_0 \text{ 被拒绝 }\right) = P_{\theta_0}\left(T \leq t_0 - 1\right) + q P_{\theta_0}\left(T = t_0\right) := \alpha.$$

因此，我们选择 t_0 满足以下条件：

$$P_{\theta_0}\left(T \leq t_0 - 1\right) \leq \alpha, \quad P_{\theta_0}\left(T \leq t_0\right) > \alpha,$$

（即 $t_0 - 1 = q_{\text{inf}}^{G}(\alpha)$，其中 q_{inf}^{G} 是第 6.1 节定义的分位数函数，而 G 是 T 的分布函数），并且

$$q = \frac{\alpha - P_{\theta_0}\left(T \leq t_0 - 1\right)}{P_{\theta_0}\left(T = t_0\right)}.$$

由于 $\phi = \phi_{\text{NP}}$ 是 Neyman-Pearson 检验，它是显著性水平 α 下的最强检验（参见第 7.1 节的 Neyman-Pearson 引理）.检验的功效为 $\beta(\theta_1)$，其中

$$\beta(\theta) := \mathbb{E}_\theta \phi(T).$$

数值例子

令 $n = 7$，则

$$\begin{aligned} P_{\theta_0}(T = 0) &= \left(\frac{1}{2}\right)^7 = 0.0078, \\ P_{\theta_0}(T = 1) &= \binom{7}{1}\left(\frac{1}{2}\right)^7 = 0.0546, \\ P_{\theta_0}(T \leq 1) &= 0.0624 > \alpha. \end{aligned}$$

因此选择 $t_0 = 1$.进一步计算得：

$$q = \frac{0.05 - 0.0078}{0.0546} = \frac{422}{546}.$$

检验的功效为：

$$\begin{aligned}\beta(\theta_1) &= P_{\theta_1}(T=0) + qP_{\theta_1}(T=1) \\ &= \left(\frac{3}{4}\right)^7 + \frac{422}{546}\binom{7}{1}\left(\frac{3}{4}\right)^6\left(\frac{1}{4}\right) \\ &= 0.1335 + \frac{422}{546}\cdot 0.3114.\end{aligned}$$

问题 2

现在我们检验：

$$H_0 : \theta_0 = \frac{1}{2},$$

对比备择假设：

$$H_1 : \theta < \frac{1}{2}.$$

在问题 1 中，构造检验 ϕ 的过程独立于 $\theta_1 < \theta_0$ 的具体值.因此，ϕ 对所有 $\theta_1 < \theta_0$ 都是最强检验（most powerful test）.我们称 ϕ 在备择假设 $H_1 : \theta < \theta_0$ 下是均匀最强检验（uniformly most powerful, UMP）.

问题 3

现在我们检验：

$$H_0 : \theta \geq \frac{1}{2},$$

对比备择假设：

$$H_1 : \theta < \frac{1}{2}.$$

回顾定义函数：

$$\beta(\theta) := \mathbb{E}_\theta \phi(T).$$

检验 ϕ 的显著性水平定义为：

$$\sup_{\theta \geq 1/2} \beta(\theta).$$

我们有：

$$\begin{aligned}\beta(\theta) &= P_\theta\left(T \leq t_0 - 1\right) + qP_\theta\left(T = t_0\right) \\ &= (1-q)P_\theta\left(T \leq t_0 - 1\right) + qP_\theta\left(T \leq t_0\right).\end{aligned}$$

注意到当 $\theta_1 < \theta_0$ 时，在 P_{θ_1} 下 T 的小值比在 P_{θ_0} 下更可能出现：

$$P_{\theta_1}(T \leq t) > P_{\theta_0}(T \leq t), \forall t \in \{0, 1, \ldots, n\}.$$

因此，$\beta(\theta)$ 是 θ 的减函数.这表明检验 ϕ 的显著性水平为：

$$\sup_{\theta \geq \frac{1}{2}} \beta(\theta) = \beta\left(\frac{1}{2}\right) = \alpha.$$

因此，ϕ 是在 $H_0 : \theta \geq \frac{1}{2}$ 对 $H_1 : \theta < \frac{1}{2}$ 的均匀最强检验.

7.3 均匀最强检验与指数族

我们现在研究 Θ 为 $\mathbb{R}$ 中区间时的情况，检验问题为：

$$H_0 : \theta \leq \theta_0,$$

对比备择假设：

$$H_1 : \theta > \theta_0.$$

假设 $\mathcal{P}$ 被某个 σ -有限测度 ν 主导.

定理 7.3.1

假设 $\mathcal{P}$ 是一维指数族：

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x)$$

并且 $c(\theta)$ 是 θ 的严格增函数.那么一个 UMP 检验 ϕ 为：

$$\phi(T(x)) := \begin{cases} 1 & \text{如果 } T(x) > t_0, \\ q & \text{如果 } T(x) = t_0, \\ 0 & \text{如果 } T(x) < t_0, \end{cases}$$

其中 q 和 t_0 满足 $\mathbb{E}_{\theta_0}\phi(T) = \alpha$.

证明：

对于 $H_0 : \theta = \theta_0$ 对 $H_1 : \theta = \theta_1$ 的 Neyman-Pearson 检验是：

$$\phi_{\text{NP}}(x) = \begin{cases} 1 & \text{如果 } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > c_0, \\ q_0 & \text{如果 } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = c_0, \\ 0 & \text{如果 } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < c_0, \end{cases}$$

其中 q_0 和 c_0 满足 $\mathbb{E}_{\theta_0}\phi_{\text{NP}}(X) = \alpha$.

我们有：

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = \exp\left[(c(\theta_1) - c(\theta_0))T(X) - (d(\theta_1) - d(\theta_0))\right].$$

因此 $\phi = \phi_{\mathrm{NP}}$.由此得出 ϕ 对 $H_0 : \theta = \theta_0$ 对 $H_1 : \theta > \theta_0$ 是 UMP.

我们接下来证明 $\beta(\theta) := \mathbb{E}_\theta \phi(T)$ 随 θ 单调递增.令

$$\bar{p}_\theta(t) = \exp[c(\theta)t - d(\theta)]$$

是 T 关于主导测度 $\bar{\nu}$ 的密度.对于 $\vartheta > \theta$，有

$$\frac{\bar{p}_\vartheta(t)}{\bar{p}_\theta(t)} = \exp[(c(\vartheta) - c(\theta))t - (d(\vartheta) - d(\theta))],$$

它是 t 的增函数.此外，有

$$\int \bar{p}_\vartheta d\bar{\nu} = \int \bar{p}_\theta d\bar{\nu} = 1.$$

因此，必定存在一个交点 s_0，使得两密度曲线交叉：

$$\begin{cases} \frac{\bar{p}_\vartheta(t)}{\bar{p}_\theta(t)} \leq 1 & \text{当 } t \leq s_0, \\ \frac{\bar{p}_\vartheta(t)}{\bar{p}_\theta(t)} \geq 1 & \text{当 } t \geq s_0. \end{cases}$$

由此可以得出：

$$\beta(\vartheta) - \beta(\theta) \geq 0.$$

因此 $\beta(\theta)$ 随 θ 递增.

最后，由于 ϕ 的显著性水平为 α，且其他显著性水平为 α 的检验 ϕ' 都满足 $\mathbb{E}_{\theta_0}\phi'(X) \leq \alpha$，因此 ϕ 是 UMP.

例子 7.3.1 正态分布方差的检验

令 $X_1, \ldots, X_n$ 是来自 $\mathcal{N}\left(\mu_0, \sigma^2\right)$ 分布的独立同分布样本，其中 μ_0 已知，而 $\sigma^2 > 0$ 是未知的.我们希望检验

$$H_0 : \sigma^2 \leq \sigma_0^2,$$

对比备择假设

$$H_1 : \sigma^2 > \sigma_0^2.$$

样本的概率密度函数为

$$\mathbf{p}_{\sigma^2}\left(x_1, \ldots, x_n\right) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu_0\right)^2 - \frac{n}{2}\log\left(2\pi\sigma^2\right)\right].$$

因此，我们可以取

$$c\left(\sigma^2\right) = -\frac{1}{2\sigma^2}$$

以及

$$T(\mathbf{X}) = \sum_{i=1}^{n} (X_i - \mu_0)^2.$$

函数 $c\left(\sigma^2\right)$ 随 σ^2 单调递增.因此我们构造检验 ϕ，其在 $T(\mathbf{X})$ 较大时拒绝 H_0.注意在 H_0 下，统计量 $T(\mathbf{X})/\sigma_0^2$ 服从自由度为 n 的 χ^2 分布，即 χ_n^2 分布（定义见第 12.2 节）.我们可以通过 χ_n^2 分布的分位数确定检验的临界值.

7.4 单侧和双侧检验：伯努利分布的例子

令 $X_1, \ldots, X_n$ 是来自 $\text{Bernoulli}(\theta)$ 分布的独立同分布样本，$0 < \theta < 1$.样本的概率密度函数为

$$\mathbf{p}_\theta\left(x_1, \ldots, x_n\right) = \exp\left[\log\left(\frac{\theta}{1-\theta}\right) \sum_{i=1}^{n} x_i + n \log(1-\theta)\right].$$

我们可以取

$$c(\theta) = \log\left(\frac{\theta}{1-\theta}\right),$$

其为 θ 的严格递增函数.此时统计量为

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$

右侧备择假设

假设

$$H_0 : \theta \leq \theta_0,$$

对比

$$H_1 : \theta > \theta_0.$$

均匀最强检验（UMP test）为

$$\phi_R(T) := \begin{cases} 1 & T > t_R, \\ q_R & T = t_R, \\ 0 & T < t_R, \end{cases}$$

其中 $\beta_R(\theta) := \mathbb{E}_\theta \phi_R(T)$ 是 θ 的严格递增函数.

左侧备择假设

假设

$$H_0 : \theta \geq \theta_0,$$

对比

$$H_1 : \theta < \theta_0.$$

均匀最强检验为

$$\phi_L(T) := \begin{cases} 1 & T < t_L, \\ q_L & T = t_L, \\ 0 & T > t_L, \end{cases}$$

其中 $\beta_L(\theta) := \mathbb{E}_\theta \phi_L(T)$ 是 θ 的严格递减函数.

双侧备择假设

假设

$$H_0 : \theta = \theta_0,$$

对比

$$H_1 : \theta \neq \theta_0.$$

检验 ϕ_R 在 $\theta > \theta_0$ 时最强，而检验 ϕ_L 在 $\theta < \theta_0$ 时最强.因此，对于双侧备择假设，均匀最强检验（UMP）不存在.

7.5 无偏检验

再次考虑一般情形：$\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ 是一族概率分布，Θ_0 和 Θ_1 是 Θ 的不相交子集，检验问题是

$$H_0 : \theta \in \Theta_0,$$

对比

$$H_1 : \theta \in \Theta_1.$$

显著性水平为 $\alpha (< 1)$.

如第 7.4 节所述，均匀最强检验（UMP）并不总是存在.因此，我们将关注一个较小的检验类，并在其中寻找均匀最强检验.

定义 7.5.1 无偏检验

若对所有 $\theta \in \Theta_0$ 和 $\vartheta \in \Theta_1$，检验 ϕ 满足：

$$\mathbb{E}_\theta \phi(X) \leq \mathbb{E}_\vartheta \phi(X),$$

则称 ϕ 为无偏检验（unbiased test）.

定义 7.5.2 UMPU 检验

若检验 ϕ 满足以下条件，则称其为均匀最强无偏检验（Uniformly Most Powerful Unbiased, UMPU）：

- ϕ 的显著性水平为 α;
- ϕ 是无偏的;
- 对于所有显著性水平为 α 的无偏检验 ϕ'，有 $\mathbb{E}_\theta\phi'(X) \leq \mathbb{E}_\theta\phi(X)\forall\theta \in \Theta_1$.

定理 7.5.1

假设 $\mathcal{P}$ 是一维指数分布族:

$$\frac{dP_\theta}{d\nu}(x) := p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x),$$

其中 $c(\theta)$ 随 θ 严格递增.那么，UMPU 检验为

$$\phi(T(x)) := \begin{cases} 1 & \text{如果 } T(x) < t_L \text{ 或 } T(x) > t_R, \\ q_L & \text{如果 } T(x) = t_L, \\ q_R & \text{如果 } T(x) = t_R, \\ 0 & \text{如果 } t_L < T(x) < t_R, \end{cases}$$

其中常数 t_R, t_L, q_R 和 q_L 需满足:

$$\mathbb{E}_{\theta_0}\phi(X) = \alpha, \quad \frac{d}{d\theta}\mathbb{E}_\theta\phi(X)\Big|_{\theta=\theta_0} = 0.$$

例子 7.5.1 正态分布均值的双侧检验

令 $X_1, \ldots, X_n$ 是来自 $\mathcal{N}\left(\mu, \sigma_0^2\right)$ 分布的独立同分布样本，其中 $\mu \in \mathbb{R}$ 未知，σ_0^2 已知.我们考虑检验

$$H_0 : \mu = \mu_0,$$

对比

$$H_1 : \mu \neq \mu_0.$$

充分统计量为 $T := \sum_{i=1}^n X_i$.对于 $t_L < t_R$，

$$\begin{aligned} \mathbb{E}_\mu\phi(T) &= \mathbb{P}_\mu\left(T > t_R\right) + \mathbb{P}_\mu\left(T < t_L\right) \\ &= \mathbb{P}_\mu\left(\frac{T - n\mu}{\sqrt{n}\sigma_0} > \frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) + \mathbb{P}_\mu\left(\frac{T - n\mu}{\sqrt{n}\sigma_0} < \frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right) \\ &= 1 - \Phi\left(\frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) + \Phi\left(\frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right), \end{aligned}$$

其中 Φ 为标准正态分布函数.

为避免与检验 ϕ 混淆，此处将标准正态密度函数记为 $\dot{\Phi}$.因此，

$$\frac{d}{d\mu}\mathbb{E}_\mu\phi(T)=\frac{n}{\sqrt{n}\sigma_0}\dot{\Phi}\left(\frac{t_R-n\mu}{\sqrt{n}\sigma_0}\right)-\frac{n}{\sqrt{n}\sigma_0}\dot{\Phi}\left(\frac{t_L-n\mu}{\sqrt{n}\sigma_0}\right).$$

令

$$\frac{d}{d\mu}\mathbb{E}_\mu\phi(T)\Big|_{\mu=\mu_0}=0,$$

得到

$$\dot{\Phi}\left(\frac{t_R-n\mu_0}{\sqrt{n}\sigma_0}\right)=\dot{\Phi}\left(\frac{t_L-n\mu_0}{\sqrt{n}\sigma_0}\right),$$

即

$$(t_R-n\mu_0)^2=(t_L-n\mu_0)^2.$$

我们取解 $(t_L-n\mu_0)=-(t_R-n\mu_0)$（因为解 $(t_L-n\mu_0)=(t_R-n\mu_0)$ 会导致检验始终拒绝，从而显著性水平 α 不成立）.代回得

$$\begin{aligned}\mathbb{E}_{\mu_0}\phi(T)&=1-\Phi\left(\frac{t_R-n\mu_0}{\sqrt{n}\sigma_0}\right)+\Phi\left(-\frac{t_R-n\mu_0}{\sqrt{n}\sigma_0}\right)\\&=2\left(1-\Phi\left(\frac{t_R-n\mu_0}{\sqrt{n}\sigma_0}\right)\right).\end{aligned}$$

要求 $\mathbb{E}_{\mu_0}\phi(T)=\alpha$，得到

$$\Phi\left(\frac{t_R-n\mu_0}{\sqrt{n}\sigma_0}\right)=1-\alpha/2,$$

因此

$$t_R-n\mu_0=\sqrt{n}\sigma_0\Phi^{-1}(1-\alpha/2),\quad t_L-n\mu_0=-\sqrt{n}\sigma_0\Phi^{-1}(1-\alpha/2).$$

7.6 条件检验 ⋆

我们现在研究参数空间 Θ 是 $\mathbb{R}^2$ 中的一个区间的情形.令 $\theta=(\beta,\gamma)$，假设 γ 是感兴趣的参数.目标是检验

$$H_0:\gamma\le\gamma_0,$$

对比备择假设

$$H_1:\gamma>\gamma_0.$$

我们进一步假设数据来自于一个指数族的标准形式：

$$p_\theta(x)=\exp\left[\beta T_1(x)+\gamma T_2(x)-d(\theta)\right]h(x).$$

在这种情况下，我们可以将检验 ϕ 限制为只依赖于充分统计量 $T=(T_1,T_2)$ 的形式.

引理 7.6.1

假设 $\{\beta:(\beta,\gamma_0)\in\Theta\}$ 包含一个开区间.定义检验函数

$$\phi(T_1,T_2)=\begin{cases}1 & \text{如果 } T_2>t_0(T_1),\\ q(T_1) & \text{如果 } T_2=t_0(T_1),\\ 0 & \text{如果 } T_2<t_0(T_1),\end{cases}$$

其中常数 $t_0(T_1)$ 和 $q(T_1)$ 允许依赖于 T_1，并满足

$$\mathbb{E}_{\gamma_0}(\phi(T_1,T_2)\mid T_1)=\alpha.$$

那么 ϕ 是均匀最强无偏检验（UMPU）.

证明思路

令 $\bar{p}_\theta(t_1,t_2)$ 表示充分统计量 (T_1,T_2) 的密度，相对于支配测度 $\bar{\nu}$:

$$\bar{p}_\theta(t_1,t_2):=\exp[\beta t_1+\gamma t_2-d(\theta)]\bar{h}(t_1,t_2).$$

假设 $\bar{\nu}(t_1,t_2)=\bar{\nu}_1(t_1)\bar{\nu}_2(t_2)$ 是一个乘积测度.T_2 在给定 $T_1=t_1$ 条件下的条件密度为

$$\begin{aligned}\bar{p}_\theta(t_2\mid t_1)&=\frac{\exp[\beta t_1+\gamma t_2-d(\theta)]\bar{h}(t_1,t_2)}{\int_{s_2}\exp[\beta t_1+\gamma s_2-d(\theta)]\bar{h}(t_1,s_2)d\bar{\nu}_2(s_2)}\\&=\exp[\gamma t_2-d(\gamma\mid t_1)]\bar{h}(t_1,t_2),\end{aligned}$$

其中

$$d(\gamma\mid t_1):=\log\left(\int_{s_2}\exp[\gamma s_2]\bar{h}(t_1,s_2)d\bar{\nu}_2(s_2)\right).$$

换句话说，T_2 在 $T_1=t_1$ 条件下的分布

- 不依赖于 β;
- 是一个参数 γ 的一维指数族分布，且具有标准形式.

因此，在给定 $T_1=t_1$ 的条件下，ϕ 是均匀最强无偏检验（UMPU）.

结果 1

检验 ϕ 的显著性水平为 α，即

$$\sup_{\gamma\le\gamma_0}\mathbb{E}_{(\beta,\gamma)}\phi(T)=\mathbb{E}_{(\beta,\gamma_0)}\phi(T)=\alpha,\forall\beta.$$

结果 1 的证明

$$\sup_{\gamma\le\gamma_0}\mathbb{E}_{(\beta,\gamma)}\phi(T)\ge\mathbb{E}_{(\beta,\gamma_0)}\phi(T)=\mathbb{E}_{(\beta,\gamma_0)}\mathbb{E}_{\gamma_0}\left(\phi(T)\mid T_1\right)=\alpha.$$

反过来，

$$\sup_{\gamma\le\gamma_0}\mathbb{E}_{(\beta,\gamma)}\phi(T)=\sup_{\gamma\le\gamma_0}\mathbb{E}_{(\beta,\gamma)}\underbrace{\mathbb{E}_{\gamma}\left(\phi(T)\mid T_1\right)}_{\le\alpha}\le\alpha.$$

结果 2

检验 ϕ 是无偏的.

结果 2 的证明

当 $\gamma>\gamma_0$ 时，条件检验无偏，因此满足 $\mathbb{E}_{\gamma}\left(\phi(T)\mid T_1\right)\ge\alpha$.因此，对于所有 β，也有

$$\mathbb{E}_{(\beta,\gamma)}\phi(T)=\mathbb{E}_{(\beta,\gamma)}\mathbb{E}_{\gamma}\left(\phi(T)\mid T_1\right)\ge\alpha,$$

即 ϕ 是无偏的.

结果 3

设 ϕ' 为一个检验，其显著性水平定义为

$$\alpha':=\sup_{\beta}\sup_{\gamma\le\gamma_0}\mathbb{E}_{(\beta,\gamma)}\phi'(T)\le\alpha,$$

并且 ϕ' 是无偏的，即满足

$$\sup_{\gamma\le\gamma_0}\sup_{\beta}\mathbb{E}_{(\beta,\gamma)}\phi'(T)\le\inf_{\gamma>\gamma_0}\inf_{\beta}\mathbb{E}_{(\beta,\gamma)}\phi'(T).$$

则在 T_1 条件下，检验 ϕ' 的显著性水平为 α'.

结果 3 的证明

由于

$$\alpha'=\sup_{\beta}\sup_{\gamma\le\gamma_0}\mathbb{E}_{(\beta,\gamma)}\phi'(T),$$

我们有

$$\mathbb{E}_{(\beta,\gamma_0)}\phi'(T)\le\alpha',\forall\beta.$$

反过来，由无偏性可知，对于所有 $\gamma>\gamma_0$，

$$\mathbb{E}_{(\beta,\gamma)}\phi'(T)\ge\alpha',\forall\beta.$$

通过连续性，可得

$$\mathbb{E}_{(\beta,\gamma_0)}\phi'(T)=\alpha',\forall\beta.$$

换句话说，

$$\mathbb{E}_{(\beta,\gamma_0)}\left(\phi'(T)-\alpha'\right)=0,\forall\beta.$$

因此也有

$$\mathbb{E}_{(\beta,\gamma_0)}\mathbb{E}_{\gamma_0}\left(\left(\phi'(T)-\alpha'\right)\mid T_1\right)=0,\forall\beta.$$

可以写成

$$\mathbb{E}_{(\beta,\gamma_0)}h\left(T_1\right)=0,\forall\beta.$$

由于假设 $\{\beta:(\beta,\gamma_0)\in\Theta\}$ 包含一个开区间，因此 T_1 对于参数 (β,γ_0) 是完备的.由此可得

$$h\left(T_1\right)=0,P_{(\beta,\gamma_0)}\text{-a.s.},\forall\beta,$$

或者由 h 的定义可得

$$\mathbb{E}_{\gamma_0}\left(\phi'(T)\mid T_1\right)=\alpha',P_{(\beta,\gamma_0)}\text{-a.s.},\forall\beta.$$

因此，在 T_1 条件下，检验 ϕ' 的显著性水平为 α'.

结果 4

设 ϕ' 是结果 3 中定义的检验，则对于任意 (β,γ)，其中 $\gamma>\gamma_0$，ϕ' 不可能比 ϕ 更有力.

结果 4 的证明

根据 Neyman-Pearson 引理，在 T_1 条件下，有

$$\mathbb{E}_{\gamma}\left(\phi'(T)\mid T_1\right)\leq\mathbb{E}_{\gamma}\left(\phi(T)\mid T_1\right),\forall\gamma>\gamma_0.$$

因此也有

$$\mathbb{E}_{(\beta,\gamma)}\phi'(T)\leq\mathbb{E}_{(\beta,\gamma)}\phi(T),\forall\beta,\gamma>\gamma_0.$$

例子 7.6.1：比较两个泊松分布的均值

考虑两个独立的样本 $\mathbf{X}=(X_1,\ldots,X_n)$ 和 $\mathbf{Y}=(Y_1,\ldots,Y_m)$，其中 $X_1,\ldots,X_n$ 是独立同分布的 $\text{Poisson}(\lambda)$，$Y_1,\ldots,Y_m$ 是独立同分布的 $\text{Poisson}(\mu)$.目标是检验

$$H_0:\lambda\leq\mu,$$

对比备择假设

$$H_1:\lambda>\mu.$$

定义

$$\beta:=\log(\mu),\gamma:=\log(\lambda/\mu).$$

该检验问题等价于

$$H_0 : \gamma \leq \gamma_0,$$

对比备择假设

$$H_1 : \gamma > \gamma_0,$$

其中 $\gamma_0 := 0$.样本的密度函数为

$$\begin{aligned}
&\mathbf{p}_\theta\left(x_1, \ldots, x_n, y_1, \ldots, y_m\right) \\
=&\exp\left[\log(\lambda)\sum_{i=1}^{n} x_i + \log(\mu)\sum_{j=1}^{m} y_j - n\lambda - m\mu\right]\prod_{i=1}^{n}\frac{1}{x_i!}\prod_{j=1}^{m}\frac{1}{y_j!} \\
=&\exp\left[\log(\mu)\left(\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} y_j\right) + \log(\lambda/\mu)\sum_{i=1}^{n} x_i - n\lambda - m\mu\right]h(\mathbf{x}, \mathbf{y}) \\
=&\exp\left[\beta T_1(\mathbf{x}, \mathbf{y}) + \gamma T_2(\mathbf{x}) - d(\theta)\right]h(\mathbf{x}, \mathbf{y})
\end{aligned}$$

其中

$$T_1(\mathbf{X}, \mathbf{Y}) := \sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j, \quad T_2(\mathbf{X}) := \sum_{i=1}^{n} X_i,$$

且

$$h(\mathbf{x}, \mathbf{y}) := \prod_{i=1}^{n}\frac{1}{x_i!}\prod_{j=1}^{m}\frac{1}{y_j!}.$$

条件分布

在给定 $T_1 = t_1$ 的条件下，T_2 的条件分布是 $\text{Binomial}\,(t_1, p)$，其中

$$p = \frac{n\lambda}{n\lambda + m\mu} = \frac{\mathrm{e}^\gamma}{1 + \mathrm{e}^\gamma}.$$

因此，在条件 $T_1 = t_1$ 下，根据从 $\text{Binomial}\,(t_1, p)$ 分布观察到的 T_2，我们检验

$$H_0 : p \leq p_0,$$

对比备择假设

$$H_1 : p > p_0,$$

其中 $p_0 := n/(n+m)$.该检验是无条件问题的 UMPU 检验.

第 8 章：估计量的比较

可以通过其风险来比较估计量.本章第 8.1 至 8.3 节定义了风险的概念，并讨论了风险的性质；第 8.4 节简要介绍了敏感性和稳健性，第 8.5 节讨论了计算方面的问题.这两节未涉及具体细节.

8.1 风险的定义

考虑一个随机变量 X，其分布为 $P_\theta, \theta \in \Theta$.令 $T = T(X)$ 为一个对感兴趣参数 $\gamma = g(\theta)$ 的估计量.一个风险函数 $R(\cdot, \cdot)$ 用于衡量因估计量误差导致的损失.风险取决于未知参数 θ 和估计量 T，定义为

$$R(\theta, T) := \mathbb{E}_\theta(L(\theta, T(X))),$$

其中 $L(\cdot, \cdot)$ 为给定的所谓损失函数.本书第 10 章将对此进行更详细的描述.

例子 8.1.1 检验的风险

考虑以下检验问题：

$$H_0 : \theta = \theta_0,$$

对比备择假设

$$H_1 : \theta = \theta_1.$$

令 $\phi(X) \in [0, 1]$ 为一个检验，则其风险定义为错误概率，即

$$R(\theta, \phi) = \begin{cases} \mathbb{E}_{\theta_0}\phi(X), & \text{若 } \theta = \theta_0, \\ 1 - \mathbb{E}_{\theta_1}\phi(X), & \text{若 } \theta = \theta_1. \end{cases}$$

例子 8.1.2 估计量的风险

当 $\gamma \in \mathbb{R}$ 时，一个重要的风险度量是均方误差（MSE）：

$$R(\theta, T) := \mathbb{E}_\theta(T(X) - g(\theta))^2 =: \mathrm{MSE}_\theta(T).$$

[2]

8.2 风险与充分性

令 $S = S(X)$ 为充分统计量.若已知充分统计量 S，则无需保留原始数据 X 即可不丢失信息.事实上，以下引理表明，基于原始数据 X 的任何决策都可以用仅依赖于 S 的随机化决策 $\delta(S)$ 替代，且其风险相同.

引理 8.2.1 假设 S 对于 θ 是充分的.令 $d : \mathcal{X} \to \mathcal{A}$ 为某一决策.则存在仅依赖于 S 的随机化决策 $\delta(S)$，使得

$$R(\theta, \delta(S)) = R(\theta, d), \forall \theta.$$

证明：令 X_s^* 为满足分布 $P(X \in \cdot \mid S = s)$ 的随机变量.由构造可知，对于所有可能的 s，X_s^* 和 X 在条件分布下相同，因此它们具有相同的分布形式.形式化地，令 Q_θ 表示 S 的分布，则

$$
\begin{aligned}
P_\theta\left(X_S^* \in \cdot\right) &= \int P\left(X_s^* \in \cdot \mid S=s\right) dQ_\theta(s) \\
&= \int P(X \in \cdot \mid S=s) dQ_\theta(s) = P_\theta(X \in \cdot).
\end{aligned}
$$

由此，取 $\delta(s) := d\left(X_s^*\right)$，引理得证.

8.3 Rao-Blackwell 定理

Rao-Blackwell 定理表明，对于凸损失，基于原始数据 X 的估计量可以用仅依赖于充分统计量 S 的估计量替代，而不会增加风险.在这种情况下不需要随机化.

引理 8.3.1 (Rao-Blackwell) 假设 S 对于 θ 是充分的.此外假设动作空间 $\mathcal{A} \subset \mathbb{R}^p$ 是凸的，并且对每个 θ，映射 $a \mapsto L(\theta, a)$ 是凸的.令 $d: \mathcal{X} \to \mathcal{A}$ 为某一决策，并定义 $d'(s) := E(d(X) \mid S=s)$（假设其存在）.则

$$
R\left(\theta, d'\right) \leq R(\theta, d), \forall \theta.
$$

证明：Jensen 不等式表明对于凸函数 g，

$$
E(g(X)) \geq g(E(X)).
$$

因此，对于所有 θ，

$$
\begin{aligned}
E(L(\theta, d(X)) \mid S=s) &\geq L(\theta, E(d(X) \mid S=s)) \\
&= L\left(\theta, d'(s)\right).
\end{aligned}
$$

根据迭代期望引理，有

$$
\begin{aligned}
R(\theta, d) &= \mathbb{E}_\theta L(\theta, d(X)) \\
&= \mathbb{E}_\theta E(L(\theta, d(X)) \mid S) \\
&\geq \mathbb{E}_\theta L\left(\theta, d'(S)\right).
\end{aligned}
$$

例子 8.3.1 均方误差

令 T 为对 $g(\theta) \in \mathbb{R}$ 的估计量，并定义

$$
R(\theta, T) := \mathbb{E}_\theta(T(X) - g(\theta))^2 =: \mathrm{MSE}_\theta(T).
$$

令 S 为充分统计量，$\tilde{T} := \mathbb{E}(T \mid S)$.根据 Rao-Blackwell 引理，

$$
R(\theta, \tilde{T}) \leq R(\theta, T), \forall \theta.
$$

均方误差可以分解为方差项和平方偏差项.由于 $\mathbb{E}\tilde{T} = \mathbb{E}T$（依据迭代期望引理），因此有

$$
\mathrm{var}_\theta(\tilde{T}) \leq \mathrm{var}_\theta(T), \forall \theta.
$$

与引理 5.2.2 和 5.3 节的 Lehmann-Scheffé 结果相比较.

8.4 敏感性与稳健性

我们可以根据对数据中大误差的敏感性比较估计量.令 $X_1, \ldots, X_n$ 为随机变量 X 的独立同分布样本.设 $T_n := T_n(X_1, \ldots, X_n)$ 为定义于每个样本量 n 的实值估计量，且在 $X_1, \ldots, X_n$ 中对称.

单个附加观测的影响

影响函数定义为

$$l(x) := (n+1)\left(T_{n+1}(X_1, \ldots, X_n, x) - T_n(X_1, \ldots, X_n)\right), \quad x \in \mathbb{R}.$$

崩溃点

令 $m \leq n$，

$$\epsilon(m) := \sup_{x_1^*, \ldots, x_m^*} |T(x_1^*, \ldots, x_m^*, X_{m+1}, \ldots, X_n)|.$$

若 $\epsilon(m) := \infty$，则称估计量在有 m 个异常值时会崩溃（break down）.崩溃点定义为

$$\epsilon^* := \min\{m : \epsilon(m) = \infty\}/n.$$

若估计量具有有界的影响函数和/或较大的崩溃点，则称其为稳健的.

8.5 计算方面的问题

如今，数据往往是高维的，参数的数量 p 也非常大.例如，最大似然估计需要对 p 个变量的函数进行最大化，当 p 很大时，这可能非常困难.如果似然函数是非凸的，或者存在一些整数值参数等问题，计算将更加复杂.此外，在第 10 章中，我们将研究贝叶斯理论，此时需要找到所谓的“后验分布”，这在计算上通常非常困难（这也是 MCMC（蒙特卡洛马尔可夫链）算法被广泛使用的原因）.显然，一个无法有效计算（例如无法在多项式时间内完成）的估计量在实际中几乎没有价值.

第 9 章

等变统计量

正如我们在第 5 章中所看到的，在某些情况下，将注意力限制在满足特定理想属性的统计量集合中可能是有用的.在第 5 章中，我们将研究范围限制在无偏估计量上.在本章中，等变性（equivariance）将是关键概念.

数据由独立同分布的实值随机变量 $X_1, \ldots, X_n$ 构成.我们记 $\mathbf{X} := (X_1, \ldots, X_n)$.相对于某种支配测度 ν 的单个观测的密度记为 p_θ，整个样本的密度为 $\mathbf{p}_\theta(\mathbf{x}) = \prod_i p_\theta(x_i)$，其中 $\mathbf{x} = (x_1, \ldots, x_n)$.

位置模型（Location Model）

在这种模型中，$\theta \in \mathbb{R}$ 是一个位置参数，假设

$$X_i = \theta + \epsilon_i, \quad i = 1, \ldots, n$$

目标是估计 θ.参数空间 Θ 和动作空间 $\mathcal{A}$ 均为实数集 $\mathbb{R}$.假设 $\epsilon_1, \ldots, \epsilon_n$ 是独立同分布的，且其密度为已知函数 $p_0(\cdot)$.

位置-尺度模型（Location-Scale Model）

在这种模型中，$\theta = (\mu, \sigma)$，其中 $\mu \in \mathbb{R}$ 是位置参数，$\sigma > 0$ 是尺度参数，假设

$$X_i = \mu + \sigma\epsilon_i, \quad i = 1, \ldots, n$$

参数空间 Θ 和动作空间 $\mathcal{A}$ 均为 $\mathbb{R} \times (0, \infty)$.假设 $\epsilon_1, \ldots, \epsilon_n$ 是独立同分布的，且其密度为已知函数 $p_0(\cdot)$.

9.1 位置模型中的等变性

定义 9.1.1 若统计量 $T = T(\mathbf{X})$ 对于任意常数 $c \in \mathbb{R}$ 和任意 $\mathbf{x} = (x_1, \ldots, x_n)$ 满足

$$T(x_1 + c, \ldots, x_n + c) = T(x_1, \ldots, x_n) + c,$$

则称 T 是位置等变统计量.

例子

$$T = \begin{cases} \bar{X} & \text{样本均值} \\ X_{(\frac{n+1}{2})} & \text{样本中位数（}n\text{ 为奇数）} \\ \cdots \end{cases}$$

定义 9.1.2 若损失函数 $L(\theta, a)$ 对于所有 $c \in \mathbb{R}$ 满足

$$L(\theta + c, a + c) = L(\theta, a), \quad (\theta, a) \in \mathbb{R}^2$$

则称 $L(\theta,a)$ 是位置不变的.

在本节中，我们将位置等变性（位置不变性）简称为等变性（不变性），并假设损失函数 $L(\theta,a)$ 是不变的.

推论 9.1.1 若统计量 T 是等变的（且 $L(\theta,a)$ 是不变的），则

$$\begin{aligned} R(\theta,T) &= \mathbb{E}_\theta L(\theta,T(\mathbf{X})) = \mathbb{E}_\theta L(0,T(\mathbf{X})-\theta) \\ &= \mathbb{E}_\theta L(0,T(\mathbf{X}-\theta)) = \mathbb{E}_\theta L_0[T(\varepsilon)] \end{aligned}$$

其中 $L_0[a] := L(0,a)$，$\varepsilon := (\epsilon_1,\dots,\epsilon_n)$.由于 ε 的分布与 θ 无关，我们可以得出风险与 θ 无关.因此可以在最后一个表达式中省略下标 θ:

$$R(\theta,T) = EL_0[T(\varepsilon)].$$

由于当 $\theta=0$ 时，$\mathbf{X}=\varepsilon$，我们还可以写成

$$R(\theta,T) = \mathbb{E}_0 L_0[T(\mathbf{X})] = R(0,T).$$

定义 9.1.3 若统计量 T 是等变的，且满足

$$R(\theta,T) = \min_{d\text{ 等变}} R(\theta,d), \quad \forall\theta,$$

或者等价地，

$$R(0,T) = \min_{d\text{ 等变}} R(0,d),$$

则称 T 为一致最小风险等变统计量（UMRE）.

9.1.1 构造 UMRE 估计量

引理 9.1.1 设 $Y_i := X_i - X_n, i=1,\dots,n$，以及 $\mathbf{Y} := (Y_1,\dots,Y_n)$.则有

$$T\text{ 是等变的 } \Leftrightarrow T(\mathbf{X}) = T(\mathbf{Y}) + X_n.$$

证明

1. ($\Rightarrow$) 显然成立.
2. ($\Leftarrow$) 将 $\mathbf{X}$ 替换为 $\mathbf{X}+c$，$\mathbf{Y}$ 保持不变（即 $\mathbf{Y}$ 是不变的）.因此，

$$T(\mathbf{X}+c) = T(\mathbf{Y}) + X_n + c = T(\mathbf{X}) + c.$$

定理 9.1.1 设 $Y_i := X_i - X_n, i = 1, \ldots, n$，$\mathbf{Y} := (Y_1, \ldots, Y_n)$，定义

$$T^*(\mathbf{Y}) := \arg\min_v E\left[L_0\left(v + \epsilon_n\right) \mid \mathbf{Y}\right].$$

此外，令

$$T^*(\mathbf{X}) := T^*(\mathbf{Y}) + X_n.$$

则 T^* 是 UMRE 估计量.

证明

1. 首先注意，$\mathbf{Y}$ 的分布不依赖于 θ，因此 T^* 确实是一个统计量.同时，由上面的引理可知 T^* 是等变的.
2. 设 T 是一个等变统计量，则有 $T(\mathbf{X}) = T(\mathbf{Y}) + X_n$.因此，

$$T(\mathbf{X}) - \theta = T(\mathbf{Y}) + \epsilon_n.$$

3. 从而，

$$R(0, T) = EL_0\left(T(\mathbf{Y}) + \epsilon_n\right) = EE\left[L_0\left(T(\mathbf{Y}) + \epsilon_n\right) \mid \mathbf{Y}\right].$$

4. 注意到

$$\begin{aligned} E\left[L_0\left(T(\mathbf{Y}) + \epsilon_n\right) \mid \mathbf{Y}\right] &\geq \min_v E\left[L_0\left(v + \epsilon_n\right) \mid \mathbf{Y}\right] \\ &= E\left[L_0\left(T^*(\mathbf{Y}) + \epsilon_n\right) \mid \mathbf{Y}\right]. \end{aligned}$$

5. 因此，

$$R(0, T) \geq EE\left[L_0\left(T^*(\mathbf{Y}) + \epsilon_n\right) \mid \mathbf{Y}\right] = R(0, T^*).$$

9.1.2 二次损失：Pitman 估计量

推论 9.1.2 若取二次损失

$$L(\theta, a) := (a - \theta)^2,$$

则有 $L_0[a] = a^2$.因此，对于 $\mathbf{Y} = \mathbf{X} - X_n$，

$$\begin{aligned} T^*(\mathbf{Y}) &= \arg\min_v E\left[\left(v + \epsilon_n\right)^2 \mid \mathbf{Y}\right] \\ &= -E\left(\epsilon_n \mid \mathbf{Y}\right). \end{aligned}$$

从而

$$T^*(\mathbf{X}) = X_n - E\left(\epsilon_n \mid \mathbf{Y}\right).$$

此估计量称为 Pitman **估计量**.

为了进一步研究二次风险的情况，我们注意到：

- 若 (X, Z) 的密度相对于 Lebesgue 测度为 $f(x, z)$，则 $Y := X - Z$ 的密度为

$$f_Y(y) = \int f(y + z, z)\, dz.$$

引理 9.1.2 设损失为二次损失，且 $\mathbf{p}_0$ 是 $\varepsilon = (\epsilon_1, \ldots, \epsilon_n)$ 相对于 Lebesgue 测度的密度，则 UMRE 统计量为

$$T^*(\mathbf{X}) = \frac{\int z\mathbf{p}_0\left(X_1 - z, \ldots, X_n - z\right) dz}{\int \mathbf{p}_0\left(X_1 - z, \ldots, X_n - z\right) dz}.$$

证明

1. 令 $\mathbf{Y} = \mathbf{X} - X_n$.随机向量 $\mathbf{Y}$ 的密度为

$$f_{\mathbf{Y}}\left(y_1, \ldots, y_{n-1}, 0\right) = \int \mathbf{p}_0\left(y_1 + z, \ldots, y_{n-1} + z, z\right) dz.$$

2. 因此，给定 $\mathbf{Y} = \mathbf{y} = (y_1, \ldots, y_{n-1}, 0)$，$\epsilon_n$ 的条件密度为

$$f_{\epsilon_n}(u) = \frac{\mathbf{p}_0\left(y_1 + u, \ldots, y_{n-1} + u, u\right)}{\int \mathbf{p}_0\left(y_1 + z, \ldots, y_{n-1} + z, z\right) dz}.$$

3. 从而

$$E\left(\epsilon_n \mid \mathbf{y}\right) = \frac{\int u\mathbf{p}_0\left(y_1 + u, \ldots, y_{n-1} + u, u\right) du}{\int \mathbf{p}_0\left(y_1 + z, \ldots, y_{n-1} + z, z\right) dz}.$$

4. 替换 $\mathbf{Y}$ 和 $\mathbf{y}$，得

$$E\left(\epsilon_n \mid \mathbf{Y}\right) = \frac{\int z\mathbf{p}_0\left(X_1 - z, \ldots, X_n - z\right) dz}{\int \mathbf{p}_0\left(X_1 - z, \ldots, X_n - z\right) dz}.$$

5. 由 $T^*(\mathbf{X}) = X_n - E\left(\epsilon_n \mid \mathbf{Y}\right)$，证毕.

例子 9.1.1 均匀分布的未知中点

设 $X_1,\dots,X_n$ 是独立同分布的 $U[\theta-1/2,\theta+1/2],\theta\in\mathbb{R}$.则

$$p_0(x)=1\{|x|\le 1/2\}.$$

可得

$$\max_{1\le i\le n}|x_i-z|\le 1/2\Leftrightarrow x_{(n)}-1/2\le z\le x_{(1)}+1/2,$$

从而

$$\mathbf{p}_0\left(x_1-z,\dots,x_n-z\right)=1\left\{x_{(n)}-1/2\le z\le x_{(1)}+1/2\right\}.$$

定义

$$T_1:=X_{(n)}-1/2,\quad T_2:=X_{(1)}+1/2,$$

则 UMRE 估计量为

$$T^*=\left(\int_{T_1}^{T_2}z\,dz\right)/\left(\int_{T_1}^{T_2}dz\right)=\frac{T_1+T_2}{2}=\frac{X_{(1)}+X_{(n)}}{2}.$$

9.1.3 不变统计量

我们现在考虑更一般的不变统计量 $\mathbf{Y}$.

定义 9.1.4 若映射 $\mathbf{Y}:\mathbb{R}^n\to\mathbb{R}^n$ 满足

$$\mathbf{Y}(\mathbf{x})=\mathbf{Y}(\mathbf{x}')\Leftrightarrow\exists c:\mathbf{x}=\mathbf{x}'+c,$$

则称其为**最大不变统计量**（maximal invariant）.
（常数 c 可依赖于 $\mathbf{x}$ 和 $\mathbf{x}'$.）

例子
映射 $\mathbf{Y}(\mathbf{x}):=\mathbf{x}-x_n$ 是最大不变统计量：

- ($\Leftarrow$) 显然成立；
- ($\Rightarrow$) 若 $\mathbf{x}-x_n=\mathbf{x}'-x_n'$，则有 $\mathbf{x}=\mathbf{x}'+(x_n-x_n')$.

更一般地，若 $d(\mathbf{X})$ 是等变的，则 $\mathbf{Y}:=\mathbf{X}-d(\mathbf{X})$ 是最大不变统计量.

定理 9.1.2
设 $d(\mathbf{X})$ 是等变的，令 $\mathbf{Y}:=\mathbf{X}-d(\mathbf{X})$，并定义

$$T^*(\mathbf{Y}):=\arg\min_v E\left[L_0(v+d(\varepsilon))\mid\mathbf{Y}\right],$$

则

$$T^*(\mathbf{X}) := T^*(\mathbf{Y}) + d(\mathbf{X})$$

是 UMRE 估计量.

证明

1. 设 T 是等变估计量，则

$$\begin{aligned} T(\mathbf{X}) &= T(\mathbf{X} - d(\mathbf{X})) + d(\mathbf{X}) \\ &= T(\mathbf{Y}) + d(\mathbf{X}). \end{aligned}$$

2. 因此

$$\begin{aligned} E\left[L_0(T(\varepsilon)) \mid \mathbf{Y}\right] &= E\left[L_0(T(\mathbf{Y}) + d(\varepsilon)) \mid \mathbf{Y}\right] \\ &\geq \min_v E\left[L_0(v + d(\varepsilon)) \mid \mathbf{Y}\right]. \end{aligned}$$

3. 使用迭代期望引理可得结论.

9.1.4 二次损失与巴苏引理（Basu's Lemma）

对于二次损失（$L_0[a] = a^2$），上述定理中 $T^*(\mathbf{Y})$ 的定义为

$$T^*(\mathbf{Y}) = -E(d(\varepsilon) \mid \mathbf{Y}) = -\mathbb{E}_0(d(\mathbf{X}) \mid \mathbf{X} - d(\mathbf{X})),$$

因此

$$T^*(\mathbf{X}) = d(\mathbf{X}) - \mathbb{E}_0(d(\mathbf{X}) \mid \mathbf{X} - d(\mathbf{X})).$$

对于等变估计量 T，有

$$T \text{ 是 UMRE } \Leftrightarrow \mathbb{E}_0(T(\mathbf{X}) \mid \mathbf{X} - T(\mathbf{X})) = 0.$$

由此可得以下结论：

1. **UMRE 估计量是无偏的**：从右侧条件可推出 $\mathbb{E}_0 T = 0$，进而 $\mathbb{E}_\theta(T) = \theta$ 对所有 θ 成立.
2. **反之亦然**：若 T 是等变且无偏的估计量，且 $T(\mathbf{X})$ 和 $\mathbf{X} - T(\mathbf{X})$ 独立，则 T 是 UMRE.

巴苏引理

设 X 的分布为 $P_\theta, \theta \in \Theta$，若 T 是充分且完全的统计量，且 $Y = Y(X)$ 的分布不依赖于 θ，则对于所有 θ，T 和 Y 在 P_θ 下独立.

证明

1. 令 A 为某可测集合，并定义

$$h(T) := P(Y \in A \mid T) - P(Y \in A).$$

2. 注意到因 T 是充分的，$P(Y \in A \mid T)$ 不依赖于 θ.由于

$$\mathbb{E}_\theta h(T) = 0, \forall \theta,$$

根据 T 的完备性，可得

$$h(T) = 0, \text{ 在 } P_\theta \text{ 上几乎处处成立 } \forall \theta,$$

即

$$P(Y \in A \mid T) = P(Y \in A), \text{ 在 } P_\theta \text{ 上几乎处处成立 } \forall \theta.$$

3. 由于 A 是任意的，因此条件分布 $P(Y \in \cdot \mid T)$ 等于无条件分布 $P(Y \in \cdot)$，从而 T 和 Y 在 P_θ 下独立.

例子 9.1.2

设 $X_1, \ldots, X_n$ 是独立同分布的 $\mathcal{N}(\theta, \sigma^2)$ 随机变量，其中 σ^2 已知.令 $T := \bar{X}$.

1. $\bar{X}$ 是充分且完全统计量；
2. $\mathbf{Y} := \mathbf{X} - \bar{X}$ 的分布不依赖于 θ.

根据巴苏引理，$\bar{X}$ 和 $\mathbf{X} - \bar{X}$ 独立，因此 $\bar{X}$ 是 UMRE 估计量.

备注

1. 即使 θ 已知或 σ^2 未知，$\bar{X}$ 和 $\mathbf{X} - \bar{X}$ 的独立性仍然成立.
2. 样本均值 $\bar{X}$ 和样本方差 $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ 独立，因为 S^2 是 $\mathbf{X} - \bar{X}$ 的函数.

9.2 在位置-尺度模型中的等变性 ⋆

位置-尺度模型

假设

$$X_i = \mu + \sigma \epsilon_i, \quad i = 1, \ldots, n$$

未知参数为 $\theta = (\mu, \sigma)$，其中 $\mu \in \mathbb{R}$ 是位置参数，$\sigma > 0$ 是尺度参数.参数空间 Θ 和作用空间 $\mathcal{A}$ 均为 $\mathbb{R} \times \mathbb{R}_+$，其中 $\mathbb{R}_+ = (0, \infty)$.假设 $\varepsilon = (\epsilon_1, \ldots, \epsilon_n)$ 的分布已知.

定义 9.2.1

若统计量 $T = T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X}))$ 对于任意常数 $b \in \mathbb{R}, c \in \mathbb{R}_+$ 和任意 $\mathbf{x} = (x_1, \ldots, x_n)$ 满足

$$T(b + cx_1, \ldots, b + cx_n) = b + cT(x_1, \ldots, x_n)$$

且

$$T_2(b+cx_1,\ldots,b+cx_n)=cT_2(x_1,\ldots,x_n),$$

则称 T 为位置-尺度等变统计量.

定义 9.2.2

若损失函数 $L(\mu,\sigma,a_1,a_2)$ 对于任意 $(\mu,a_1,b)\in\mathbb{R}^3,(\sigma,a_2,c)\in\mathbb{R}_+^3$ 满足

$$L(b+c\mu,c\sigma,b+ca_1,ca_2)=L(\mu,\sigma,a_1,a_2),$$

则称 L 为位置-尺度不变损失函数.

在本节中，位置-尺度等变（不变）简记为等变（不变），并假设损失函数 $L(\theta,a)$ 是不变的.

推论 9.2.1

若 T 是等变的（且 $L(\theta,a)$ 是不变的），则

$$\begin{aligned}R(\theta,T)&=\mathbb{E}_\theta L(\mu,\sigma,T_1(\mathbf{X}),T_2(\mathbf{X}))\\&=\mathbb{E}_\theta L\left(0,1,\frac{T_1(\mathbf{X})-\mu}{\sigma},\frac{T_2(\mathbf{X})}{\sigma}\right)\\&=\mathbb{E}_\theta L\left(0,1,T_1(\varepsilon),T_2(\varepsilon)\right)=\mathbb{E}_\theta L_0(T(\varepsilon))\end{aligned}$$

其中 $L_0(a_1,a_2):=L(0,1,a_1,a_2)$.由此可知风险不依赖于 θ.因此在最后的表达式中可以省略下标 θ:

$$R(\theta,T)=EL_0(T(\varepsilon))$$

定义 9.2.3

若统计量 T 满足

$$R(\theta,T)=\min_{d\text{ 等变}}R(\theta,d),\quad\forall\theta,$$

或等价地，

$$R(0,1,T_1,T_2)=\min_{d\text{ 等变}}R(0,1,d_1,d_2),$$

则称 T 为统一最小风险等变（UMRE）统计量.

9.2.1 构造 UMRE 估计量 ⋆

定理 9.2.1

假设 $d(\mathbf{X})$ 是等变的.令

$$\mathbf{Y} := \frac{\mathbf{X} - d_1(\mathbf{X})}{d_2(\mathbf{X})}$$

和

$$T^*(\mathbf{Y}) := \arg \min_{a_1 \in \mathbb{R}, a_2 \in \mathbb{R}_+} E\left[L_0(d_1(\varepsilon) + d_2(\varepsilon)a_1, d_2(\varepsilon)a_2) \mid \mathbf{Y}\right],$$

则

$$T^*(\mathbf{X}) := \begin{pmatrix} d_1(\mathbf{X}) + d_2(\mathbf{X})T_1^*(\mathbf{Y}) \\ d_2(\mathbf{X})T_2^*(\mathbf{Y}) \end{pmatrix}$$

是 UMRE 估计量.

证明

1. 由等变性可得：

$$\mathbf{Y} = \frac{\mathbf{X} - d_1(\mathbf{X})}{d_2(\mathbf{X})} = \frac{\varepsilon - d_1(\varepsilon)}{d_2(\varepsilon)},$$

因此

$$\varepsilon = d_1(\varepsilon) + d_2(\varepsilon)\mathbf{Y}.$$

2. 设 T 是等变估计量，则

$$\begin{aligned} EL_0(T_1(\varepsilon), T_2(\varepsilon)) &= EL_0(T_1(d_1(\varepsilon) + d_2(\varepsilon)\mathbf{Y}), T_2(d_1(\varepsilon) + d_2(\varepsilon)\mathbf{Y})) \\ &= EL_0(d_1(\varepsilon) + d_2(\varepsilon)T_1(\mathbf{Y}), d_2(\varepsilon)T_2(\mathbf{Y})) \\ &= EE\left[L_0(d_1(\varepsilon) + d_2(\varepsilon)T_1(\mathbf{Y}), d_2(\varepsilon)T_2(\mathbf{Y})) \mid \mathbf{Y}\right] \\ &\geq E \min_{a_1 \in \mathbb{R}, a_2 \in \mathbb{R}_+} E\left[L_0(d_1(\varepsilon) + d_2(\varepsilon)a_1, d_2(\varepsilon)a_2) \mid \mathbf{Y}\right] \\ &= EE\left[L_0(d_1(\varepsilon) + d_2(\varepsilon)T_1^*(\mathbf{Y}), d_2(\varepsilon)T_2^*(\mathbf{Y})) \mid \mathbf{Y}\right]. \end{aligned}$$

由此得证.

9.2.2 二次损失 ⋆

对于二次损失函数 $\left(L_0\left(a_1, a_2\right) := a_1^2\right)$，上节定理中 $T^*(\mathbf{Y})$ 的定义为

$$T^*(\mathbf{Y}) = \arg\min_{a_1 \in \mathbb{R}} E\left[\left(d_1(\varepsilon) + d_2(\varepsilon)a_1\right)^2 \mid \mathbf{Y}\right].$$

于是，我们有以下结论：

引理 9.2.1

假设 d 是等变的，且充分且完全，则

$$T^*(\mathbf{X}) := d_1(\mathbf{X}) - d_2(\mathbf{X})\frac{Ed_1(\varepsilon)d_2(\varepsilon)}{Ed_2^2(\varepsilon)}$$

是 UMRE.

证明

根据 Basu 引理，d 和 $\mathbf{Y}$ 是独立的.因此

$$E\left[(d_1(\varepsilon)+d_2(\varepsilon)a_1)^2 \mid \mathbf{Y}\right] = E(d_1(\varepsilon)+d_2(\varepsilon)a_1)^2.$$

此外，

$$\arg\min_{a_1\in\mathbb{R}} E(d_1(\varepsilon)+d_2(\varepsilon)a_1)^2 = -\frac{Ed_1(\varepsilon)d_2(\varepsilon)}{Ed_2^2(\varepsilon)}.$$

由此得证.

例 9.2.1

正态分布均值的 UMRE：未知 σ^2

设 $X_1,\ldots,X_n$ 是独立同分布的，且服从 $\mathcal{N}(\mu,\sigma^2)$ 分布.定义

$$d_1(\mathbf{X}) := \bar{X}, \quad d_2(\mathbf{X}) := S,$$

其中 S^2 是样本方差：

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2.$$

显然，d 是等变的.此外，由例 4.3.6 可知，d 是充分的，应用引理 5.4.1 可知 d 也是完全的.此外，

$$Ed_1(\varepsilon) = E\bar{\epsilon} = 0,$$

并且根据例 9.1.2（Basu 引理的结果），我们知道 $d_1(\mathbf{X}) = \bar{X}$ 和 $d_2(\mathbf{X}) = S$ 是独立的.因此，

$$Ed_1(\varepsilon)d_2(\varepsilon) = Ed_1(\varepsilon)Ed_2(\varepsilon) = 0.$$

由此可得，$T^*(\mathbf{X}) = \bar{X}$ 是 UMRE.

第十章

决策理论

在本章中，观测随机变量（即数据）记为 $X \in \mathcal{X}$，其分布为 $P \in \mathcal{P}$.概率模型为 $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$，其中 θ 是未知参数.

在一些特定情况下，我们将 X 替换为向量 $\mathbf{X} = (X_1, \ldots, X_n)$，其中 $X_1, \ldots, X_n$ 是独立同分布的，其分布为 $P \in \{P_\theta : \theta \in \Theta\}$（因此 $\mathbf{X}$ 的分布为 $\mathbb{P} := \prod_{i=1}^n P \in \left\{\mathbb{P}_\theta = \prod_{i=1}^n P_\theta : \theta \in \Theta\right\}$）.

10.1 决策及其风险

我们在第八章中定义了风险函数，这里将更为正式地阐述这一概念.令 $\mathcal{A}$ 为行动空间.

- 当 $\mathcal{A} = \mathbb{R}$ 时，表示估计一个实值参数.
- 当 $\mathcal{A} = \{0, 1\}$ 时，表示检验假设.
- 当 $\mathcal{A} = [0, 1]$ 时，表示随机化检验.
- 当 $\mathcal{A} = \{\text{区间}\}$ 时，表示置信区间.

给定观测值 X，我们决定采取 $\mathcal{A}$ 中的某一行动.因此，决策是一个映射 $d : \mathcal{X} \to \mathcal{A}$，其中 $d(X)$ 为采取的行动.如果 $\mathcal{A} = \mathbb{R}$，通常称决策为估计量（如记为 T）.如果 $\mathcal{A} = \{0, 1\}$ 或 $\mathcal{A} = [0, 1]$，通常称决策为检验（如记为 ϕ）.

损失函数（Verlustfunktion） 是一个映射：

$$L : \Theta \times \mathcal{A} \to \mathbb{R},$$

其中 $L(\theta, a)$ 表示参数值为 θ 时，采取行动 a 所导致的损失.

决策 $d(X)$ 的风险定义为：

$$R(\theta, d) := \mathbb{E}_\theta L(\theta, d(X)), \quad \theta \in \Theta.$$

例 10.1.1 参数估计
在估计感兴趣的参数 $g(\theta) \in \mathbb{R}$ 的情况下，行动空间为 $\mathcal{A} = \mathbb{R}$（或其子集）.常见的损失函数包括：

$$L(\theta, a) := w(\theta)|g(\theta) - a|^r,$$

其中 $w(\cdot)$ 是给定的非负权重，$r \geq 0$ 是给定的幂次.相应的风险为：

$$R(\theta, d) = w(\theta)\mathbb{E}_\theta|g(\theta) - d(X)|^r.$$

一个特例是取 $w \equiv 1$ 且 $r = 2$.此时，损失函数为 $L(\theta, a) = (g(\theta) - a)^2$，称为二次损失，而

$$R(\theta, d) = \mathbb{E}_\theta|g(\theta) - d(X)|^2$$

称为均方误差.

例 10.1.2 检验问题

考虑检验假设：

$$H_0 : \theta \in \Theta_0,$$

与备择假设：

$$H_1 : \theta \in \Theta_1.$$

这里，Θ_0 和 Θ_1 是给定的参数空间子集，且 $\Theta_0 \cap \Theta_1 = \emptyset$.行动空间为 $\mathcal{A} = \{0, 1\}$，损失函数为：

$$L(\theta, a) := \begin{cases} 1 & \text{若 } \theta \in \Theta_0 \text{ 且 } a = 1, \\ c & \text{若 } \theta \in \Theta_1 \text{ 且 } a = 0, \\ 0 & \text{其他情况}, \end{cases}$$

其中 $c > 0$ 是给定常数.相应的风险为：

$$R(\theta, d) = \begin{cases} P_\theta(d(X) = 1) & \text{若 } \theta \in \Theta_0, \\ cP_\theta(d(X) = 0) & \text{若 } \theta \in \Theta_1, \\ 0 & \text{其他情况}. \end{cases}$$

因此，风险对应于第一类错误和第二类错误的概率.

注:

最优决策 d 是风险 $R(\theta, d)$ 最小的决策.然而，由于 θ 是未知的，比较两个决策函数 d_1 和 d_2 时可能出现问题：$R(\theta, d_1)$ 在某些 θ 值下可能小于 $R(\theta, d_2)$，而在其他 θ 值下却可能大于 $R(\theta, d_2)$.

例 10.1.3 均值估计

我们回顾例 5.2.1.设 $X \in \mathbb{R}$，令 $g(\theta) = \mathbb{E}_\theta X := \mu$，并取二次损失函数：

$$L(\theta, a) := |\mu - a|^2.$$

假设 $\mathrm{var}_\theta(X) = 1$ 对所有 θ 都成立.考虑以下决策集合：

$$d_\lambda(X) := \lambda X, \quad 0 \le \lambda \le 1.$$

其风险为：

$$\begin{aligned} R(\theta, d_\lambda) &= \mathrm{var}(\lambda X) + \mathrm{bias}_\theta^2(\lambda X) \\ &= \lambda^2 + (\lambda - 1)^2 \mu^2. \end{aligned}$$

"最优" 的 λ 值为：

$$\lambda_{\mathrm{opt}} := \frac{\mu^2}{1 + \mu^2},$$

因为该值最小化了 $R(\theta, d_\lambda)$.然而，λ_{opt} 依赖于未知参数 μ，因此 $d_{\lambda_{\mathrm{opt}}}(X)$ 不能作为估计量.

各种最优性概念

我们将讨论三种最优性概念：可容性（Admissibility, Zulässigkeit）、极小极大（Minimax）和贝叶斯最优性（Bayes）.

10.2 可容性决策

定义 10.2.1

当满足以下条件时，决策 d' 被称为**严格优于** d：

$$R\left(\theta, d'\right) \leq R(\theta, d), \quad \forall \theta,$$

且

$$\exists \theta: \quad R\left(\theta, d'\right) < R(\theta, d).$$

当存在 d' 严格优于 d 时，称 d 是**不可容的（inadmissible）**.

例 10.2.1 仅利用一个观测值

设 $n \geq 2$，$X_1, \ldots, X_n$ 是独立同分布的随机变量，且 $g(\theta) := \mathbb{E}_\theta\left(X_i\right) := \mu$，$\operatorname{var}\left(X_i\right) = 1$（对所有 i）.采用二次损失 $L(\theta, a) := |\mu - a|^2$.考虑以下两个决策：

1. $d'\left(X_1, \ldots, X_n\right) := \bar{X}_n$，即样本均值；
2. $d\left(X_1, \ldots, X_n\right) := X_1$，即第一个样本值.

则对于任意 θ：

$$R\left(\theta, d'\right) = \frac{1}{n}, \quad R(\theta, d) = 1.$$

因此，d 是不可容的.

注意：

- 要证明某个决策 d 是不可容的，只需找到一个严格优于它的 d' 即可.
- 然而，要证明某个决策 d 是可容的，需要验证不存在任何严格优于它的 d'.这通常需要考虑所有可能的 d'.

10.2.1 不使用数据的决策是可容的

令 $L(\theta, a) := |g(\theta) - a|^r$，并令 $d(X) := g(\theta_0)$，其中 θ_0 是某个固定值.

引理 10.2.1
假设 P_{θ_0} 主导（dominates）P_θ 对所有 θ 都成立，则 d 是可容的.

证明：

假设 d' 优于 d，则有：

$$\mathbb{E}_{\theta_0} \left| g\left(\theta_0\right) - d'(X) \right|^r \leq 0.$$

这意味着：

$$d'(X) = g\left(\theta_0\right), \quad P_{\theta_0}\text{-几乎处处成立} \tag{10.1}$$

由 (10.1) 可知：

$$P_{\theta_0}\left(d'(X) \neq g\left(\theta_0\right)\right) = 0.$$

假设 P_{θ_0} 主导 P_θ，则有：

$$P_\theta\left(d'(X) \neq g\left(\theta_0\right)\right) = 0, \quad \forall \theta.$$

因此，对于所有 θ，都有 $d'(X) = g\left(\theta_0\right)$, P_θ -几乎处处成立，从而：

$$R\left(\theta, d'\right) = R(\theta, d).$$

所以，d' 并未严格优于 d，从而 d 是可容的.

10.2.2 Neyman-Pearson 检验是可容的

考虑检验问题：

$$H_0 : \theta = \theta_0,$$

与备择假设：

$$H_1 : \theta = \theta_1.$$

定义检验 ϕ 的风险 $R(\theta, \phi)$ 为第一类错误和第二类错误的概率：

$$R(\theta, \phi) := \begin{cases} \mathbb{E}_\theta \phi(X), & \theta = \theta_0, \\ 1 - \mathbb{E}_\theta \phi(X), & \theta = \theta_1. \end{cases}$$

令 p_0 和 p_1 分别为 P_{θ_0} 和 P_{θ_1} 相对于某个主导测度 ν 的密度（例如 $\nu = P_{\theta_0} + P_{\theta_1}$）.一个 Neyman-Pearson 检验定义为（见第 7.1 节）：

$$\phi_{\mathrm{NP}} := \begin{cases} 1, & \text{若 } p_1/p_0 > c, \\ q, & \text{若 } p_1/p_0 = c, \\ 0, & \text{若 } p_1/p_0 < c, \end{cases}$$

其中 $0 \le q \le 1$，$0 \le c < \infty$ 是给定的常数.

引理 10.2.2

Neyman-Pearson 检验是可容的，当且仅当以下两种情况之一成立：

1. 其检验功效严格小于 1；
2. 它在所有功效为 1 的检验中具有最小的显著性水平.

证明：

假设 $R(\theta_0, \phi) < R(\theta_0, \phi_{\mathrm{NP}})$.根据 Neyman-Pearson 引理，我们知道要么 $R(\theta_1, \phi) > R(\theta_1, \phi_{\mathrm{NP}})$（即 ϕ 并不优于 ϕ_{NP}），要么 $c = 0$.但当 $c = 0$ 时，有 $R(\theta_1, \phi_{\mathrm{NP}}) = 0$，即 ϕ_{NP} 的功效为 1.

类似地，若假设 $R(\theta_1, \phi) < R(\theta_1, \phi_{\mathrm{NP}})$，则由 Neyman-Pearson 引理可得 $R(\theta_0, \phi) > R(\theta_0, \phi_{\mathrm{NP}})$，因为假设 $c < \infty$.

10.3 极小极大决策

定义 10.3.1

一个决策 d 被称为极小极大（minimax）的，如果满足

$$\sup_{\theta} R(\theta, d) = \inf_{d'} \sup_{\theta} R\left(\theta, d'\right)$$

也就是说，极小极大准则关注的是在最坏情况下表现最好的决策.

10.3.1 极小极大 Neyman-Pearson 检验

引理 10.3.1

Neyman-Pearson 检验 ϕ_{NP} 是极小极大的，当且仅当

$$R(\theta_0, \phi_{\mathrm{NP}}) = R(\theta_1, \phi_{\mathrm{NP}}).$$

证明：

令 ϕ 为一个检验，并令 $j = 0, 1$ 时，

$$r_j := R(\theta_j, \phi_{\mathrm{NP}}), \quad r'_j := R(\theta_j, \phi).$$

假设 $r_0 = r_1$ 且 ϕ_{NP} 不是极小极大的，则对于某个检验 ϕ，

$$\max_j r'_j < \max_j r_j.$$

这意味着同时存在：

$$r_0' < r_0, \quad r_1' < r_1.$$

根据 Neyman-Pearson 引理，这是不可能的.
令 $S = \{(R(\theta_0, \phi), R(\theta_1, \phi)) : \phi : \mathcal{X} \to [0,1]\}$，注意到 S 是凸集.如果 $r_0 < r_1$，我们可以找到一个检验 ϕ，使得 $r_0 < r_0' < r_1$ 且 $r_1' < r_1$.因此，ϕ_{NP} 不是极小极大的.同理，若 $r_0 > r_1$ 时亦然.

10.4 贝叶斯决策

假设参数空间 Θ 是一个可测空间，并赋予其概率测度 Π.我们称 Π 为**先验分布（a priori distribution）**.

定义 10.4.1
贝叶斯风险（Bayes risk）相对于概率测度 Π 定义为

$$r(\Pi, d) := \int_\Theta R(\vartheta, d) d\Pi(\vartheta).$$

一个决策 d 被称为相对于 Π 的贝叶斯最优（Bayes optimal），如果

$$r(\Pi, d) = \inf_{d'} r\left(\Pi, d'\right).$$

假设 Π 相对于某个主导测度 μ 有密度 $w := d\Pi/d\mu$，则可以写为

$$r(\Pi, d) = \int_\Theta R(\vartheta, d) w(\vartheta) d\mu(\vartheta) := r_w(d).$$

因此，贝叶斯风险可以理解为风险的加权平均.例如，可以对“重要”的 θ 值赋予更高的权重.

10.4.1 贝叶斯检验

重新考虑检验问题：

$$H_0 : \theta = \theta_0,$$

与备择假设：

$$H_1 : \theta = \theta_1.$$

令损失函数为：

$$L(\theta_0, a) := a, \quad L(\theta_1, a) := 1 - a,$$

先验权重为：

$$w(\theta_0) =: w_0, \quad w(\theta_1) =: w_1 = 1 - w_0.$$

则贝叶斯风险为：

$$r_w(\phi) := w_0 R(\theta_0, \phi) + w_1 R(\theta_1, \phi).$$

取 $0 < w_0 = 1 - w_1 < 1$.

引理 10.4.1
贝叶斯检验为：

$$\phi_{\text{Bayes}} = \begin{cases} 1, & \text{若 } p_1/p_0 > w_0/w_1, \\ q, & \text{若 } p_1/p_0 = w_0/w_1, \\ 0, & \text{若 } p_1/p_0 < w_0/w_1. \end{cases}$$

证明：

$$\begin{aligned} r_w(\phi) &= w_0 \int \phi p_0 + w_1 \left(1 - \int \phi p_1\right) \\ &= \int \phi (w_0 p_0 - w_1 p_1) + w_1. \end{aligned}$$

选择 $\phi \in [0,1]$ 来最小化 $\phi(w_0 p_0 - w_1 p_1)$，最优选择为：

$$\phi = \begin{cases} 1, & \text{若 } w_0 p_0 - w_1 p_1 < 0, \\ q, & \text{若 } w_0 p_0 - w_1 p_1 = 0, \\ 0, & \text{若 } w_0 p_0 - w_1 p_1 > 0, \end{cases}$$

其中 q 可以取任意 0 和 1 之间的值.

注意：

$$2r_w(\phi_{\text{Bayes}}) = 1 - \int |w_1 p_1 - w_0 p_0|.$$

特别地，当 $w_0 = w_1 = 1/2$ 时，

$$2r_w(\phi_{\text{Bayes}}) = 1 - \int |p_1 - p_0|/2,$$

即，当两个密度函数接近时，风险较大.

10.5 贝叶斯估计的构造

设 X 的分布为 $P \in \mathcal{P} := \{P_\theta : \theta \in \Theta\}$.假设 $\mathcal{P}$ 被一个（σ-有限的）测度 ν 主导，且 $p_\theta = dP_\theta/d\nu$ 为密度函数.令 Π 为 Θ 上的先验分布，其密度为 $w := d\Pi/d\mu$.我们将 p_θ 理解为 θ 已知时 X 的条件密度，记为：

$$p_\theta(x) = p(x \mid \theta), \quad x \in \mathcal{X}.$$

此外，定义边际密度为：

$$p(\cdot) := \int_{\Theta} p(\cdot \mid \vartheta) w(\vartheta) d\mu(\vartheta).$$

定义 10.5.1

θ 的后验密度为：

$$w(\vartheta \mid x) = p(x \mid \vartheta) \frac{w(\vartheta)}{p(x)}, \quad \vartheta \in \Theta,\ x \in \mathcal{X}.$$

引理 10.5.1

给定数据 $X = x$，将 θ 视为具有密度 $w(\vartheta \mid x)$ 的随机变量.定义：

$$l(x, a) := E[L(\theta, a) \mid X = x] = \int_{\Theta} L(\vartheta, a) w(\vartheta \mid x) d\mu(\vartheta),$$

$$d(x) := \arg\min_{a} l(x, a).$$

则 d 为贝叶斯决策 d_{Bayes}.

证明：

$$\begin{aligned}
r_w\left(d'\right) &= \int_{\Theta} R\left(\vartheta, d'\right) w(\vartheta) d\mu(\vartheta) \\
&= \int_{\Theta} \left[\int_{\mathcal{X}} L\left(\vartheta, d'(x)\right) p(x \mid \vartheta) d\nu(x) \right] w(\vartheta) d\mu(\vartheta) \\
&= \int_{\mathcal{X}} \left[\int_{\Theta} L\left(\vartheta, d'(x)\right) w(\vartheta \mid x) d\mu(\vartheta) \right] p(x) d\nu(x) \\
&= \int_{\mathcal{X}} l\left(x, d'(x)\right) p(x) d\nu(x) \\
&\geq \int_{\mathcal{X}} l(x, d(x)) p(x) d\nu(x) \\
&= r_w(d).
\end{aligned}$$

推论 10.5.1

贝叶斯决策为：

$$d_{\text{Bayes}}(X) = \arg\min_{a \in \mathcal{A}} l(X, a),$$

其中：

$$\begin{aligned}
l(x, a) = E(L(\theta, a) \mid X = x) &= \int L(\vartheta, a) w(\vartheta \mid x) d\mu(\vartheta) \\
&= \int L(\vartheta, a) g_{\vartheta}(S(x)) w(\vartheta) d\mu(\vartheta) h(x) / p(x).
\end{aligned}$$

因此：

$$d_{\text{Bayes}}(X) = \arg\min_{a\in\mathcal{A}} \int L(\vartheta, a) g_\vartheta(S) w(\vartheta) d\mu(\vartheta),$$

这仅依赖于充分统计量 S.

10.5.1 贝叶斯检验的重新审视

考虑检验问题：

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1,$$

损失函数为：

$$L(\theta_0, a) := a, \quad L(\theta_1, a) := 1 - a, \quad a \in \{0, 1\}.$$

则：

$$l(x, \phi) = \phi w_0 p_0(x)/p(x) + (1-\phi) w_1 p_1(x)/p(x).$$

因此：

$$\arg\min_{\phi} l(\cdot, \phi) = \begin{cases} 1, & \text{若 } w_1 p_1 > w_0 p_0, \\ q, & \text{若 } w_1 p_1 = w_0 p_0, \\ 0, & \text{若 } w_1 p_1 < w_0 p_0. \end{cases}$$

10.5.2 贝叶斯估计量的二次损失情况

在以下结果中，我们将用到：

引理 10.5.2
设 Z 为实值随机变量，则：

$$\arg\min_{a\in\mathbb{R}} E(Z-a)^2 = EZ.$$

证明：

$$E(Z-a)^2 = \operatorname{var}(Z) + (a - EZ)^2.$$

对于 $\mathcal{A} = \mathbb{R}$ 且 $\Theta \subseteq \mathbb{R}$，若 $L(\theta, a) := |\theta - a|^2$，则：

$$d_{\text{Bayes}}(X) = E(\theta \mid X).$$

对于二次损失，假设 $T = E(\theta \mid X)$，则一个估计量 T' 的贝叶斯风险为：

$$r_w\left(T'\right)=E\operatorname{var}(\theta\mid X)+E\left(T-T'\right)^2,$$

（参见引理 5.2.2）.这一结论可通过如下直接计算得出：

$$\begin{aligned}r_w\left(T'\right)&=\int R\left(\vartheta,T'\right)w(\vartheta)d\mu(\vartheta)\\&=ER\left(\theta,T'\right)=E\left(\theta-T'\right)^2\\&=E\left[E\left((\theta-T')^2\mid X\right)\right],\end{aligned}$$

其中 θ 是随机变量：

$$\begin{aligned}E\left((\theta-T')^2\mid X\right)&=E\left((\theta-T)^2\mid X\right)+\left(T-T'\right)^2\\&=\operatorname{var}(\theta\mid X)+\left(T-T'\right)^2.\end{aligned}$$

10.5.3 贝叶斯估计量与最大后验估计量（MAP）

再次考虑 $\Theta\subseteq\mathbb{R}$ 的情形，且 $\mathcal{A}=\Theta$，损失函数为 $L(\theta,a):=1\{|\theta-a|>c\}$，其中 $c>0$ 是给定的常数.则：

$$l(x,a)=\Pi(|\theta-a|>c\mid X=x)=\int_{|\vartheta-a|>c}w(\vartheta\mid x)d\vartheta.$$

当 $c\to 0$ 时，有：

$$\frac{1-l(x,a)}{2c}=\frac{\Pi(|\theta-a|\le c\mid X=x)}{2c}\approx w(a\mid x)=p(x\mid a)\frac{w(a)}{p(x)}.$$

因此，当 c 较小时，贝叶斯规则近似为：

$$d_0(x):=\arg\max_{a\in\Theta}p(x\mid a)w(a).$$

估计量 $d_0(X)$ 称为**最大后验估计量（MAP）**.如果 w 是 Θ 上的均匀分布密度（仅当 Θ 有界时存在），则 $d_0(X)$ 为最大似然估计量.

10.5.4 三个具体例子

在许多例子中，后验分布 $w(\vartheta\mid x)$ 的完整表达式不必展开，这样可以节省大量工作.首先关注其如何依赖于 ϑ，其他部分可以在之后通过积分等理论手段（尽管可能计算上复杂）获得.我们回顾符号 $\propto$（见 4.9 节）.例如，可写为：

$$\begin{aligned}w(\vartheta\mid x)&=p(x\mid\vartheta)w(\vartheta)/p(x)\\&\propto p(x\mid\vartheta)w(\vartheta),\end{aligned}$$

因为边际密度 $p(x)$ 不依赖于 ϑ.

在以下三个例子中，后验分布与先验分布属于同一分布族.我们称此类先验分布为对应分布的**共轭先验分布**.

例子 10.5.1 带伽马先验的泊松分布

假设在给定 θ 的情况下，X 服从参数为 θ 的泊松分布，且 θ 服从 $\mathrm{Gamma}(k,\lambda)$ 分布.θ 的密度为：

$$w(\vartheta)=\lambda^k\vartheta^{k-1}\mathrm{e}^{-\lambda\vartheta}/\Gamma(k),$$

其中：

$$\Gamma(k)=\int_0^\infty \mathrm{e}^{-z}z^{k-1}dz.$$

$\mathrm{Gamma}(k,\lambda)$ 分布的均值为：

$$E\theta=\int_0^\infty \vartheta w(\vartheta)d\vartheta=\frac{k}{\lambda}.$$

后验密度为：

$$\begin{aligned}w(\vartheta\mid x)&=p(x\mid\vartheta)\frac{w(\vartheta)}{p(x)}\\&=\mathrm{e}^{-\vartheta}\frac{\vartheta^x}{x!}\frac{\lambda^k\vartheta^{k-1}\mathrm{e}^{-\lambda\vartheta}/\Gamma(k)}{p(x)}\\&=\mathrm{e}^{-\vartheta(1+\lambda)}\vartheta^{k+x-1}c(x,k,\lambda),\end{aligned}$$

其中 $c(x,k,\lambda)$ 使得：

$$\int w(\vartheta\mid x)d\vartheta=1.$$

使用 $\propto$ 符号表示：

$$\begin{aligned}w(\vartheta\mid x)&\propto p(x\mid\vartheta)w(\vartheta)\\&\propto\mathrm{e}^{-\vartheta(1+\lambda)}\vartheta^{k+x-1}.\end{aligned}$$

可以看出，这节省了大量书写.我们识别出 $w(\vartheta\mid x)$ 是 $\mathrm{Gamma}(k+x,1+\lambda)$ 分布的密度.

- 带二次损失的贝叶斯估计量为：

$$E(\theta\mid X)=\frac{k+X}{1+\lambda}.$$

- 最大后验估计量为：

$$\frac{k+X-1}{1+\lambda}.$$

例子 10.5.2 带 Beta 先验的二项分布

假设在给定 θ 的情况下，X 服从 $\text{Binomial}(n,\theta)$ 分布，且 θ 在 $[0,1]$ 上均匀分布.则：

$$w(\vartheta \mid x) = \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}/p(x).$$

这是 $\text{Beta}(x+1, n-x+1)$ 分布的密度.因此，带二次损失的贝叶斯估计量为：

$$E(\theta \mid X) = \frac{X+1}{n+2}.$$

更一般地，假设 $X \sim \text{Binomial}(n,\theta)$ 且 θ 具有 $\text{Beta}(r,s)$ 先验分布：

$$w(\vartheta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\vartheta^{r-1}(1-\vartheta)^{s-1}, \quad 0 < \vartheta < 1,$$

其中 r 和 s 为给定的正数.先验期望为：

$$E\theta = \frac{r}{r+s}.$$

带二次损失的贝叶斯估计量为后验期望：

$$E(\theta \mid X) = \frac{X+r}{n+r+s}.$$

10.5.3 正态分布的贝叶斯估计与最大后验估计量（MAP）

假设 $X \sim \mathcal{N}(\theta, 1)$ 且 $\theta \sim \mathcal{N}\left(c, \tau^2\right)$，其中 c 和 τ^2 是已知常数.后验密度为：

$$\begin{aligned} w(\vartheta \mid x) &= \frac{p(x \mid \vartheta)w(\vartheta)}{p(x)} \\ &\propto \phi(x-\vartheta)\phi\left(\frac{\vartheta - c}{\tau}\right) \\ &\propto \exp\left[-\frac{1}{2}\left\{(x-\vartheta)^2 + \frac{(\vartheta-c)^2}{\tau^2}\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\vartheta - \frac{\tau^2 x + c}{\tau^2+1}\right\}^2 \frac{1+\tau^2}{\tau^2}\right]. \end{aligned}$$

因此，带二次损失的贝叶斯估计为：

$$T_{\text{Bayes}} = E(\theta \mid X) = \frac{\tau^2 X + c}{\tau^2 + 1}.$$

10.6 贝叶斯方法的讨论

主观性问题

对贝叶斯方法的主要反对意见在于它通常是主观的.最终的估计量对先验分布的选择非常敏感.然而，贝叶斯方法非常强大，且在许多情况下显得自然.如果先验是从之前的数据集中获得

的，那么上述的主观性问题会变得不那么显著.此外，在具有多个未知参数的复杂模型中，贝叶斯方法是开发合理算法的重要工具.

可信区间

参数的（频率论意义的）置信区间可能难以找到，而且对于“非专家”来说难以解释.贝叶斯版本的置信区间称为**可信区间（credibility set）**，一般被认为是一个更直观的概念.例如，对于实值参数 θ，$(1-\alpha)$ 可信区间定义为：

$$I := \left[\hat{\theta}_L(X), \hat{\theta}_R(X)\right],$$

其中端点 $\hat{\theta}_L$ 和 $\hat{\theta}_R$ 满足：

$$\int_{\hat{\theta}_L(X)}^{\hat{\theta}_R(X)} w(\vartheta \mid X) d\vartheta = (1-\alpha).$$

也就是说，这是一个后验概率为 $(1-\alpha)$ 的区间.然而，一个 $(1-\alpha)$ 可信区间通常不是 $(1-\alpha)$ 置信区间，即从频率论的角度看，其性质并不总是明确的.

实用观点

贝叶斯方法对构造估计量非常有用.我们可以随后研究贝叶斯方法的频率论性质.例如，对于 $\text{Binomial}(n, \theta)$ 模型，假设 θ 具有均匀先验分布，则贝叶斯估计为：

$$\hat{\theta}_{\text{Bayes}}(X) = \frac{X+1}{n+2}.$$

给定此估计量，我们可以“忘记”它是通过贝叶斯方法得到的，并研究其（频率论）均方误差等性质.

复杂度正则化

以下是一个“玩具”例子，说明贝叶斯方法如何帮助构造有用的估计方法.假设 $X_1, \ldots, X_n$ 是独立随机变量，其中 $X_i \sim \mathcal{N}(\theta_i, 1)$，且 n 个参数 θ_i 均未知.这是观测数与未知参数数目相等的情况，需要复杂度正则化.

复杂度正则化（见第 16 章）意味着原则上允许任何参数值，但对于选择“复杂”值会付出代价.“复杂性”的含义依具体情况而定.在此例中，“复杂性”是稀疏性的对立面，稀疏性定义为向量 ϑ 非零元素的个数.考虑如下估计量：

$$\hat{\theta} := \arg\min_{\vartheta \in \mathbb{R}^n} \sum_{i=1}^{n} (X_i - \vartheta_i)^2 + 2\lambda \sum_{i=1}^{n} |\vartheta_i|,$$

其中 $\lambda > 0$ 是正则化参数.注意，当 $\lambda = 0$ 时，有 $\hat{\theta}_i = X_i$；而当 $\lambda = \infty$ 时，有 $\hat{\theta} \equiv 0$.随着 λ 增大，估计量越稀疏.事实上，可以验证，对于 $i = 1, \ldots, n$：

$$\hat{\theta}_i = \begin{cases} X_i - \lambda & X_i > \lambda \\ 0 & |X_i| \le \lambda \\ X_i + \lambda & X_i < -\lambda \end{cases}.$$

这称为**软阈值估计量（soft thresholding estimator）**.该方法对应于带有双指数（也称拉普拉斯）先验的贝叶斯最大后验估计.假设先验为 $\theta_1, \ldots, \theta_n$ 独立同分布，其密度为：

$$w(z) = \frac{1}{\tau\sqrt{2}} \exp\left[-\frac{\sqrt{2}|z|}{\tau}\right], \quad z \in \mathbb{R},$$

其中 $\tau > 0$ 为先验的尺度参数（τ^2 为该分布的方差）.给定 $X_1, \ldots, X_n$，向量 θ 的后验分布为：

$$w\left(\vartheta \mid X_1, \ldots, X_n\right) \propto (2\pi)^{-n/2} \exp\left[-\frac{\sum_{i=1}^n \left(X_i - \vartheta_i\right)^2}{2}\right] \times (2\pi\tau)^{-n/2} \exp\left[-\frac{\sqrt{2}\sum_{i=1}^n |\vartheta_i|}{\tau}\right].$$

因此，带有正则化参数 $\lambda = \sqrt{2}/\tau$ 的 $\hat{\theta}$ 是最大后验估计量.

贝叶斯方法作为理论工具

在第 11 章中，我们将说明贝叶斯方法可以作为证明频率论下界等结果的工具.例如，具有常数风险的贝叶斯估计量也是极小极大估计量.此类结果的核心思想是寻找“最坏的先验分布”.

10.7 积分去除参数 ⋆

为了构造灵活的先验分布，可以将其建模为依赖于另一个“超参数” τ，即：

$$w(\vartheta) := w(\vartheta \mid \tau).$$

在固定 τ 的情况下，将 ϑ 积分去除后，X 的密度为：

$$\tilde{p}(x \mid \tau) := \int p(x \mid \vartheta) w(\vartheta \mid \tau) d\mu(\vartheta).$$

可以通过例如最大似然法（通常计算较为困难）或矩方法估计 τ，从而得到基于估计参数 $\hat{\tau}$ 的先验 $w(\vartheta \mid \hat{\tau})$.这种基于数据的先验称为**经验贝叶斯（empirical Bayes）**方法.

例 10.7.1：具有超参数的伽马先验泊松分布

假设 $X_1, \ldots, X_n$ 相互独立，且 X_i 服从泊松分布 $\mathrm{Poisson}(\theta_i)$，其中 $i = 1, \ldots, n$.假设 $\theta_1, \ldots, \theta_n$ 是独立同分布的，且服从 $\mathrm{Gamma}(k, \lambda)$ 分布，即每个的先验密度为：

$$w(z \mid k, \lambda) = \mathrm{e}^{-\lambda z} z^{k-1} \lambda^k / \Gamma(k), \quad z > 0.$$

此处 k 和 λ 被视为超参数.则 $X_1, \ldots, X_n$ 的联合密度为：

$$\begin{aligned}
&\tilde{\mathbf{p}}\left(x_1, \ldots, x_n \mid k, \lambda\right) \\
&\propto \int\left(\mathrm{e}^{-\sum_{i=1}^n \vartheta_i} \prod_{i=1}^n \vartheta_i^{x_i} \mathrm{e}^{-\lambda \sum_{i=1}^n \vartheta_i} \prod_{i=1}^n \vartheta_i^{k-1} \frac{\lambda^k}{\Gamma(k)}\right) d\vartheta_1 \cdots d\vartheta_n \\
&=\prod_{i=1}^n \frac{\Gamma\left(x_i+k\right)}{\Gamma(k)} p^k(1-p)^{x_i+k-1},
\end{aligned}$$

其中 $p := \lambda/(1+\lambda)$.因此，在 $\tilde{\mathbf{p}}(\cdot \mid k, \lambda)$ 下，观测值 $X_1, \ldots, X_n$ 是独立的，且 X_i 服从负二项分布，其参数为 k 和 p（可参考负二项分布公式，见例如例 2.4.1）.通过直接计算或查阅教材可得负二项分布的均值和方差：

$$E\left(X_i \mid k, \lambda\right)=\frac{k(1-p)}{p}=\frac{k}{\lambda},$$

$$\operatorname{var}\left(X_i \mid k, \lambda\right)=\frac{k(1-p)}{p^2}=\frac{k(1+\lambda)}{\lambda^2}.$$

我们使用矩方法估计 k 和 λ.令 $\bar{X}_n$ 为样本均值，$S_n^2 := \sum_{i=1}^n \left(X_i-\bar{X}_n\right)^2/(n-1)$ 为样本方差.解以下方程：

$$\frac{\hat{k}}{\hat{\lambda}}=\bar{X}_n, \quad \frac{\hat{k}(1+\hat{\lambda})}{\hat{\lambda}^2}=S_n^2,$$

得到：

$$\hat{k}=\frac{\bar{X}_n^2}{S_n^2-\bar{X}_n}, \quad \hat{\lambda}=\frac{\bar{X}_n}{S_n^2-\bar{X}_n}.$$

对于给定的 k 和 λ，θ_i 的贝叶斯估计如例 10.5.1 中所示.将估计值 $\hat{k}$ 和 $\hat{\lambda}$ 代入后得到经验贝叶斯估计量：

$$\hat{\theta}_i=\frac{X_i+\hat{k}}{1+\hat{\lambda}}=X_i\left(1-\frac{\bar{X}_n}{S_n^2}\right)+\frac{\bar{X}_n^2}{S_n^2}, \quad i=1, \ldots, n.$$

X_i 本身是 θ_i 的最大似然估计值（MLE）.可以看出，经验贝叶斯估计量 $\hat{\theta}_i$ 使用所有观测值来估计特定的 θ_i.经验贝叶斯估计量 $\hat{\theta}_i$ 是 X_i 和 $\bar{X}_n$ 的凸组合：

$$\hat{\theta}_i=(1-\alpha) X_i+\alpha \bar{X}_n,$$

其中 $\alpha=\bar{X}_n/S_n^2$.如果总体样本的均值和方差大致相等（即总体样本接近泊松分布），则 α 通常接近 1.

第十一章

证明可容性与极小极大性

贝叶斯估计对于顽固的频率派研究者也非常有用.它们可以用来构造极小极大（minimax）或可容（admissible）的估计量，或者验证某个估计量是否具有极小极大性或可容性.

回顾基本定义.设 $X \in \mathcal{X}$ 的分布为 $P_\theta, \theta \in \Theta$，$T = T(X)$ 是一个统计量（估计或决策），$L(\theta, a)$ 是一个损失函数，$R(\theta, T) := \mathbb{E}_\theta L(\theta, T(X))$ 是 T 的风险.
$-T$ 是极小极大的，如果对所有 T' 满足 $\sup_\theta R(\theta, T) \leq \sup_\theta R(\theta, T')$.
$\circ T$ 是不可容的，如果存在 T' 满足 $\forall \theta, R(\theta, T') \leq R(\theta, T)$ 且 $\exists \theta, R(\theta, T') < R(\theta, T)$.

- 如果对某先验密度 w，T 满足 $\forall T', r_w(T) \leq r_w(T')$，则称 T 为贝叶斯（Bayes）估计.

贝叶斯风险 $r_w(T)$ 定义为：

$$r_w(T) = \int R(\vartheta, T) w(\vartheta) d\mu(\vartheta)$$

如果我们说一个统计量 T 是贝叶斯估计，但未指定具体的先验分布，则表示存在某个先验使得 T 是贝叶斯估计.如果 $R(\theta, T) = R(T)$ 与 θ 无关，则 T 的贝叶斯风险也与先验无关.

对于无法使用均匀分布作为先验的情况（例如 Θ 是无界的），扩展贝叶斯（extended Bayes）的概念非常有用.

定义 11.0.1
统计量 T 被称为扩展贝叶斯估计（extended Bayes），如果存在一列先验密度 $\{w_m\}_{m=1}^\infty$（其主导测度允许依赖于 m），满足：

$$r_{w_m}(T) - \inf_{T'} r_{w_m}(T') \to 0, \quad m \to \infty.$$

11.1 极小极大性

引理 11.1.1
若统计量 T 的风险 $R(\theta, T) = R(T)$ 与 θ 无关，则：
(i) T 可容 $\Rightarrow$ T 极小极大，
(ii) T 贝叶斯 $\Rightarrow$ T 极小极大，更一般地，
(iii) T 扩展贝叶斯 $\Rightarrow$ T 极小极大.

证明：

(i) 若 T 可容，则对所有 T'，要么存在某个 θ 使得 $R(\theta, T') > R(T)$，要么 $\forall \theta, R(\theta, T') \geq R(T)$.因此，$\sup_\theta R(\theta, T') \geq R(T)$.

(ii) 贝叶斯性蕴含扩展贝叶斯性，因此由 (iii) 得证.然而，我们单独给出一个简单证明.
首先，对任意 T'，有：

$$r_w\left(T'\right)=\int R\left(\vartheta,T'\right)w(\vartheta)d\mu(\vartheta)$$

由如下不等式可得：

$$\begin{aligned}r_w\left(T'\right)&\leq\int\sup_{\vartheta}R\left(\vartheta,T'\right)w(\vartheta)d\mu(\vartheta)\\&=\sup_{\vartheta}R\left(\vartheta,T'\right).\end{aligned}$$

即，贝叶斯风险总是被上确界风险所界定.假设存在统计量 T'，使得 $\sup_\theta R\left(\theta,T'\right)<R(T)$，则有：

$$r_w\left(T'\right)\leq\sup_{\vartheta}R\left(\vartheta,T'\right)<R(T)=r_w(T),$$

这与 T 是贝叶斯估计的假设矛盾.

(iii) 为简便起见，假设对于每个 m，存在先验 w_m 对应的贝叶斯决策 T_m，满足：

$$r_{w_m}\left(T_m\right)=\inf_{T'}r_{w_m}\left(T'\right),\quad m=1,2,\ldots$$

由假设，对于任意 $\epsilon>0$，存在足够大的 m，使得：

$$R(T)=r_{w_m}(T)\leq r_{w_m}\left(T_m\right)+\epsilon\leq\sup_{\theta}R\left(\theta,T'\right)+\epsilon,$$

由于贝叶斯风险被上确界风险界定（见 11.1），并且 ϵ 可以任意小，(iii) 得证.

例 11.1.1
二项分布的极小极大估计量

设 $X\sim\text{Binomial}(n,\theta)$，$\theta\in(0,1)$ 的先验分布为 $\text{Beta}(r,s)$.则对于平方损失，贝叶斯估计量为：

$$T=\frac{X+r}{n+r+s},$$

其风险为：

$$R(\theta,T)=\mathbb{E}_\theta(T-\theta)^2,$$

计算得：

$$\begin{aligned}R(\theta,T)&=\text{var}_\theta(T)+\text{bias}_\theta^2(T)\\&=\frac{n\theta(1-\theta)}{(n+r+s)^2}+\left[\frac{n\theta+r}{n+r+s}-\theta\right]^2\\&=\frac{\left[(r+s)^2-n\right]\theta^2+\left[n-2r(r+s)\right]\theta+r^2}{(n+r+s)^2}.\end{aligned}$$

风险为常数，当且仅当 θ^2 和 θ 的系数为零：

$$(r+s)^2-n=0,\quad n-2r(r+s)=0,$$

解得：

$$r=s=\sqrt{n}/2.$$

将其代入估计量得：

$$T=\frac{X+\sqrt{n}/2}{n+\sqrt{n}}$$

是极小极大的.对应的极小极大风险为：

$$R(T)=\frac{1}{4(\sqrt{n}+1)^2}.$$

与无偏估计量 X/n 的上确界风险比较：

$$\sup_\theta R\left(\theta,X/n\right)=\sup_\theta\frac{\theta(1-\theta)}{n}=\frac{1}{4n},$$

可以看出当 n 较大时，这两者相差不大.

11.1.1 Pitman 估计量的极小极大性 ⋆

我们再次讨论 Pitman 估计量（参见引理 9.1.2）：

$$T^*=\frac{\int z\mathbf{p}_0\left(X_1-z,\ldots,X_n-z\right)dz}{\int\mathbf{p}_0\left(X_1-z,\ldots,X_n-z\right)dz}$$

引理 11.1.2

对于平方损失，T^* 是扩展贝叶斯估计.

证明：

设 w_m 是区间 $[-m,m]$ 上均匀分布的密度：

$$w_m=1_{[-m,m]}/2m.$$

其后验密度为：

$$w_m(\vartheta\mid x)=\frac{p_0(x-\vartheta)1_{[-m,m]}(\vartheta)}{\int_{-m}^m p_0(x-\vartheta)d\vartheta}.$$

贝叶斯估计量为：

$$T_m=\frac{\int_{-m}^m\vartheta p_0(x-\vartheta)d\vartheta}{\int_{-m}^m p_0(x-\vartheta)d\vartheta}.$$

我们现在计算 $R\left(\theta,T_m\right)=\mathbb{E}_\theta(T_m-\theta)^2$.令：

$$T_{a,b}(x):=\frac{\int_a^b zp_0(x-z)dz}{\int_a^b p_0(x-z)dz}.$$

对于所有 x，当 $a \to -\infty$ 且 $b \to \infty$ 时，$T_{a,b}(x) \to T(x)$.同时易验证：

$$\lim_{a\to-\infty,b\to\infty} \mathbb{E}_0 T_{a,b}^2(X) \to \mathbb{E}_0 T^2(X).$$

（注意：对于任何先验 w，$\mathbb{E}_0 T^2(X)$ 是贝叶斯风险 $r_w(T)$，因为 $R(\theta,T) = \mathbb{E}_0 T^2(X)$ 与 θ 无关.）此外，

$$T_{a,b}(X) - \theta = \frac{\int_a^b (z-\theta) p_0(X-z)dz}{\int_a^b p_0(x-z)dz} = \frac{\int_{a-\theta}^{b-\theta} v p_0(X-\theta-v)dv}{\int_{a-\theta}^{b-\theta} p_0(X-\theta-v)dv}.$$

由此可得：

$$\mathbb{E}_\theta (T_{a,b}(X) - \theta)^2 = \mathbb{E}_0 T_{a-\theta,b-\theta}^2(X).$$

因此：

$$R(\theta, T_m) = \mathbb{E}_0 T_{-m-\theta,m-\theta}^2(X).$$

贝叶斯风险为：

$$r_{w_m}(T_m) = \mathbb{E}_{\theta\sim w_m} R(\theta, T_m) = \frac{1}{2m}\int_{-m}^{m} \mathbb{E}_0 T_{-m-\vartheta,m-\vartheta}^2(X) d\vartheta.$$

因此，对于任意 $0 < \epsilon < 1$：

$$\begin{aligned} r_{w_m}(T_m) &\geq \inf_{|\vartheta|\leq m(1-\epsilon)} (1-\epsilon)\mathbb{E}_0 T_{-m-\vartheta,m-\vartheta}^2(X) \\ &\geq \inf_{a\leq -m\epsilon, b\geq m\epsilon} (1-\epsilon)\mathbb{E}_0 T_{a,b}^2(X). \end{aligned}$$

由此，对于任意 $0 < \epsilon < 1$：

$$\liminf_{m\to\infty} r_{w_m}(T_m) \geq \liminf_{m\to\infty} \inf_{a\leq -m\epsilon, b\geq m\epsilon} (1-\epsilon)\mathbb{E}_0 T_{a,b}^2(X) = (1-\epsilon)\mathbb{E}_0 T^2(X).$$

因此，$r_{w_m}(T_m) \to \mathbb{E}_0 T^2(X)$，即 $r_{w_m}(T_m) - r_{w_m}(T) \to 0$.

推论 11.1.1

T^* 是极小极大的（对于平方损失）.

11.2 可容许性

在本节中，假设参数空间是一个拓扑空间的开子集，从而可以讨论 Θ 中元素的开邻域以及定义在 Θ 上的连续函数.此外，我们仅考虑满足 $R(\theta,T) < \infty$ 的统计量 T.

引理 11.2.1

设统计量 T 是先验密度 w 的贝叶斯决策，则以下条件之一为真时，T 是可容许的：
(i) T 是唯一的贝叶斯决策，即若 $r_w(T) = r_w(T')$，则 $\forall\theta, T = T'$ P_θ -几乎处处成立；
(ii) 对任意 T'，$R(\theta,T')$ 关于 θ 是连续的，并且对任意开集 $U \subset \Theta$，U 的先验概率 $\Pi(U) := \int_U w(\vartheta)d\mu(\vartheta)$ 严格为正.

证明：

(i) 假设存在某个 T'，使得 $\forall\theta, R\left(\theta, T'\right) \leq R(\theta, T)$.则 $r_w\left(T'\right) \leq r_w(T)$.由于 T 是贝叶斯决策，我们必须有

$$r_w\left(T'\right) = r_w(T).$$

因此，$\forall\theta, T'$ 和 T 在 P_θ -几乎处处相等，从而 $\forall\theta, R\left(\theta, T'\right) = R(\theta, T)$，即 T' 不可能比 T 严格更优.

(ii) 假设 T 是不可容许的，则存在某个 T' 使得 $\forall\theta, R\left(\theta, T'\right) \leq R(\theta, T)$ 且对某个 $\theta_0, R\left(\theta_0, T'\right) < R\left(\theta_0, T\right)$.因此存在 $\epsilon > 0$ 和 θ_0 的开邻域 $U \subset \Theta$，使得

$$R\left(\vartheta, T'\right) \leq R(\vartheta, T) - \epsilon, \forall\vartheta \in U.$$

于是，

$$\begin{aligned} r_w\left(T'\right) &= \int_U R\left(\vartheta, T'\right) w(\vartheta) d\nu(\vartheta) + \int_{U^c} R\left(\vartheta, T'\right) w(\vartheta) d\nu(\vartheta) \\ &\leq \int_U R(\vartheta, T) w(\vartheta) d\nu(\vartheta) - \epsilon\Pi(U) + \int_{U^c} R(\vartheta, T) w(\vartheta) d\nu(\vartheta) \\ &= r_w(T) - \epsilon\Pi(U) < r_w(T). \end{aligned}$$

这与 T 是贝叶斯决策相矛盾.

引理 11.2.2

设 T 是扩展贝叶斯统计量，且对任意 T'，$R\left(\theta, T'\right)$ 关于 θ 是连续的.此外，假设对于任意开集 $U \subset \Theta$，

$$\frac{r_{w_m}(T) - \inf_{T'} r_{w_m}\left(T'\right)}{\Pi_m(U)} \to 0,$$

其中 $\Pi_m(U) := \int_U w_m(\vartheta) d\mu_m(\vartheta)$ 是先验 Π_m 下 U 的概率.当 $m \to \infty$ 时，T 是可容许的.

证明：

假设 T 是不可容许的，则存在某个 T' 使得 $\forall\theta, R\left(\theta, T'\right) \leq R(\theta, T)$ 且对某个 $\theta_0, R\left(\theta_0, T'\right) < R\left(\theta_0, T\right)$，因此存在 $\epsilon > 0$ 和 θ_0 的开邻域 $U \subset \Theta$，使得

$$R\left(\vartheta, T'\right) \leq R(\vartheta, T) - \epsilon, \forall\vartheta \in U.$$

这意味着对所有 m，

$$r_{w_m}\left(T'\right) \leq r_{w_m}(T) - \epsilon\Pi_m(U).$$

假设 T_m 是先验 w_m 的贝叶斯决策，则

$$r_{w_m}\left(T_m\right) = \inf_{T'} r_{w_m}\left(T'\right), \forall m.$$

因此，对于所有 m，

$$r_{w_m}(T_m) \le r_{w_m}(T') \le r_{w_m}(T) - \epsilon \Pi_m(U),$$

即，

$$\frac{r_{w_m}(T) - r_{w_m}(T_m)}{\Pi_m(U)} \ge \epsilon > 0.$$

这与假设相矛盾.

11.2.1 正态分布均值的可容许估计

设 $X \sim \mathcal{N}(\theta, 1)$，$\theta \in \Theta := \mathbb{R}$，且平方风险为 $R(\theta, T) := \mathbb{E}_\theta (T - \theta)^2$.考虑以下形式的估计量：

$$T = aX + b, \quad a > 0,\, b \in \mathbb{R}.$$

引理

T 是可容许的，当且仅当以下任一情况成立：

(i) $a < 1$，

(ii) $a = 1$ 且 $b = 0$.

证明

$(\Leftarrow)$(i)

首先，我们需要证明 T 是某个先验分布下的贝叶斯估计量.令先验分布为正态分布，即 $\theta \sim \mathcal{N}\left(c, \tau^2\right)$，其中 c 和 τ^2 的具体值将稍后确定.在例子 10.5.3 中，我们已经得知贝叶斯估计量为：

$$T_{\text{Bayes}} = E(\theta \mid X) = \frac{\tau^2 X + c}{\tau^2 + 1}$$

令

$$\frac{\tau^2}{\tau^2 + 1} = a, \quad \frac{c}{\tau^2 + 1} = b,$$

于是 $T = T_{\text{Bayes}}$.

接下来，我们验证引理 11.2.1 的条件 (i)，即 T 是唯一的贝叶斯估计量.根据 10.5.2 的推导，

$$r_w\left(T'\right) = E \operatorname{var}(\theta \mid X) + E\left(T - T'\right)^2,$$

若 $r_w\left(T'\right) = r_w(T)$，则

$$E\left(T - T'\right)^2 = 0.$$

这里，期望是对 θ 积分的结果，即关于具有密度

$$p(x) = \int p_\vartheta(x) w(\vartheta) d\mu(\vartheta)$$

的测度 P.

我们可以写作 $X=\theta+\epsilon$，其中 $\theta\sim\mathcal{N}\left(c,\tau^2\right)$，$\epsilon$ 是独立的标准正态随机变量.由此，$X\sim\mathcal{N}\left(c,\tau^2+1\right)$，即 P 是 $\mathcal{N}\left(c,\tau^2+1\right)$ 分布.

由于 $E(T-T')^2=0$，可得 $T=T'$ P -几乎处处成立.因为 P 支配所有 P_θ，所以对所有 θ，$T=T'$ P_θ -几乎处处成立.因此 T 是唯一的，从而 T 是可容许的.

$(\Leftarrow)(\mathrm{ii})$
此时，$T=X$.我们使用引理 11.2.2.由于 $R(\theta,T)=1$ 对所有 θ 成立，因此对于任意先验分布，$r_w(T)=1$.

设 w_m 是 $\mathcal{N}(0,m)$ 分布的密度.在例子 10.5.3 和证明的第一部分中，我们已经知道贝叶斯估计量为：

$$T_m=\frac{m}{m+1}X.$$

利用偏差-方差分解，其风险为：

$$R\left(\theta,T_m\right)=\frac{m^2}{(m+1)^2}+\left(\frac{m}{m+1}-1\right)^2\theta^2=\frac{m^2}{(m+1)^2}+\frac{\theta^2}{(m+1)^2}.$$

因为 $E\theta^2=m$，所以贝叶斯风险为：

$$r_{w_m}\left(T_m\right)=\frac{m^2}{(m+1)^2}+\frac{m}{(m+1)^2}=\frac{m}{m+1}.$$

因此，

$$r_{w_m}(T)-r_{w_m}\left(T_m\right)=1-\frac{m}{m+1}=\frac{1}{m+1}.$$

这表明 T 是扩展贝叶斯统计量.

现在，我们验证引理 11.2.2 的更强性质.对于开区间 $U=(u,u+h)$，其中 u 和 $h>0$ 固定，有：

$$\begin{aligned}\Pi_m(U)&=\Phi\left(\frac{u+h}{\sqrt{m}}\right)-\Phi\left(\frac{u}{\sqrt{m}}\right)\\&=\frac{1}{\sqrt{m}}\phi\left(\frac{u}{\sqrt{m}}\right)h+o(1/\sqrt{m}).\end{aligned}$$

当 m 足够大时，

$$\phi\left(\frac{u}{\sqrt{m}}\right)\approx\phi(0)=\frac{1}{\sqrt{2\pi}}>\frac{1}{4},$$

因此，

$$\Pi_m(U) \geq \frac{1}{4\sqrt{m}} h.$$

于是，

$$\frac{r_{w_m}(T) - r_{w_m}(T_m)}{\Pi_m(U)} \leq \frac{4}{h\sqrt{m}}.$$

右侧在 $m \to \infty$ 时收敛到 0，这表明 X 是可容许的.

$(\Rightarrow)$
我们现在需要证明，若 (i) 或 (ii) 不成立，则 T 是不可容许的.

1. $a > 1$:
   此时，

$$R(\theta, aX + b) \geq \operatorname{var}(aX + b) > 1 = R(\theta, X)，$$

所以 $aX + b$ 是不可容许的.

2. $a = 1, b \neq 0$:
   此时，偏差项导致不可容许性：

$$R(\theta, X + b) = 1 + b^2 > 1 = R(\theta, X).$$

11.3 指数族分布中可容许估计量 ⋆

引理 11.3.1
令 $\theta \in \Theta = \mathbb{R}$，且 $\{P_\theta : \theta \in \Theta\}$ 为标准形式的指数族分布：

$$p_\theta(x) = \exp[\theta T(x) - d(\theta)] h(x).$$

在二次损失下（即损失函数 $L(\theta, a) := |a - g(\theta)|^2$)），$T$ 是 $g(\theta) := \dot{d}(\theta)$ 的可容许估计量.

证明：
回忆以下性质：

$$\dot{d}(\theta) = \mathbb{E}_\theta T, \quad \ddot{d}(\theta) = \operatorname{var}_\theta(T) = I(\theta)$$

(参见第 4.8 节).令 T' 为一个估计量，其期望为：

$$\mathbb{E}_\theta T' := q(\theta).$$

T' 的偏差为：

$$b(\theta) = q(\theta) - g(\theta)，$$

即：

$$q(\theta) = b(\theta) + g(\theta) = b(\theta) + \dot{d}(\theta).$$

这意味着：

$$\dot{q}(\theta) = \dot{b}(\theta) + I(\theta).$$

根据 Cramer-Rao 下界：

$$\begin{aligned} R\left(\theta, T'\right) &= \operatorname{var}_\theta\left(T'\right) + b^2(\theta) \\ &\geq \frac{[\dot{q}(\theta)]^2}{I(\theta)} + b^2(\theta) = \frac{[\dot{b}(\theta) + I(\theta)]^2}{I(\theta)} + b^2(\theta). \end{aligned}$$

假设：

$$R\left(\theta, T'\right) \leq R(\theta, T), \quad \forall \theta,$$

由于 $R(\theta, T) = I(\theta)$，这意味着：

$$\frac{[\dot{b}(\theta) + I(\theta)]^2}{I(\theta)} + b^2(\theta) \leq I(\theta),$$

或等价于：

$$I(\theta)\left\{b^2(\theta) + 2\dot{b}(\theta)\right\} \leq -[\dot{b}(\theta)]^2 \leq 0.$$

从而得到：

$$b^2(\theta) + 2\dot{b}(\theta) \leq 0,$$

因此 $b(\theta)$ 是递减的.如果 $b(\theta) \neq 0$，则：

$$\frac{\dot{b}(\theta)}{b^2(\theta)} \leq -\frac{1}{2},$$

从而：

$$\frac{d}{d\theta}\left(\frac{1}{b(\theta)}\right) - \frac{1}{2} \geq 0,$$

即：

$$\frac{d}{d\theta}\left(\frac{1}{b(\theta)} - \frac{\theta}{2}\right) \geq 0.$$

换句话说，$\frac{1}{b(\theta)} - \frac{\theta}{2}$ 是一个递增函数.

我们现在证明这将导致矛盾，从而 $b(\theta) = 0$ 对所有 θ 成立.

反证法：

假设 $b\left(\theta_0\right)<0$ 对某些 θ_0 成立，则对于所有 $\vartheta>\theta_0$，因为 $b(\cdot)$ 是递减的，有：

$$b(\vartheta)<0.$$

因此：

$$\frac{1}{b(\vartheta)} \geq \frac{1}{b\left(\theta_0\right)}+\frac{\vartheta-\theta_0}{2} \rightarrow \infty, \quad \vartheta \rightarrow \infty,$$

即：

$$b(\vartheta) \rightarrow 0, \quad \vartheta \rightarrow \infty.$$

这与 $b(\theta)$ 是递减函数相矛盾.

类似地，如果 $b\left(\theta_0\right)>0$，取 $\vartheta \rightarrow-\infty$，同样可以推导出矛盾.

因此 $b(\theta)=0$ 对所有 θ 成立，即 T' 是 θ 的无偏估计量.根据 Cramer-Rao 下界，可以得到：

$$R\left(\theta, T'\right)=\operatorname{var}_\theta\left(T'\right) \geq R(\theta, T)=I(\theta).$$

例子 11.3.1

正态分布中均值的样本均值是可容许的估计量

令 $X \sim \mathcal{N}(\theta, 1)$，$\theta \in \mathbb{R}$.此时，$X$ 是 θ 的可容许估计量.

例子 11.3.2

方差未知情况下正态分布的不可容许性

令 $X \sim \mathcal{N}\left(0, \sigma^2\right)$，$\sigma^2 \in(0, \infty)$.密度为：

$$p_\theta(x)=\exp[\theta T(x)-d(\theta)] h(x),$$

其中：

$$T(x)=-x^2 / 2, \quad \theta=1 / \sigma^2, \quad d(\theta)=-\frac{\log \theta}{2}.$$

我们定义 $T_a=-a X^2$，其中 $T=T_{1 / 2}$.其风险为：

$$R\left(\theta, T_a\right)=2 a^2 \sigma^4+\left(a-\frac{1}{2}\right)^2 \sigma^4.$$

当 $a=1 / 6$ 时风险最小：

$$R\left(\theta, T_{1 / 6}\right)=\frac{\sigma^4}{6}<\frac{\sigma^4}{2}=R(\theta, T).$$

因此，T 是不可容许的.

11.4 高维情况下的不可容许性 ⋆

假设 $X_i \sim \mathcal{N}(\theta_i, 1)$，$i = 1, \ldots, p$，且 $X_1, \ldots, X_p$ 相互独立.向量 $\theta := (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$ 是未知的.对于估计量 $T = (T_1, \ldots, T_p) \in \mathbb{R}^p$，定义风险函数为：

$$R(\theta, T) := \sum_{i=1}^{p} \mathbb{E}_\theta (T_i - \theta_i)^2.$$

注意，对于样本均值估计量 $X := (X_1, \ldots, X_p)$，其风险为 $R(\theta, X) = p$.通过类似于 $p = 1$ 的分析，可以证明 X 是极小极大的、扩展贝叶斯的、UMRE（统一最小风险等变估计量），且达到了 Cramer-Rao 下界.然而，当 $p > 2$ 时，X 是不可容许的.这一点可以通过以下引理得出：X 可以通过 Stein 估计量改进.

我们使用以下记号：$\|X\|^2 := \sum_{i=1}^{p} X_i^2$.

定义 11.4.1

令 $p > 2$ 且 $0 < b < 2(p-2)$ 为某个常数，Stein 估计量定义为：

$$T^* := \left(1 - \frac{b}{\|X\|^2}\right) X.$$

引理 11.4.1

Stein 估计量的风险为：

$$R\left(\theta, T^*\right) = p - \left[2b(p-2) - b^2\right] \mathbb{E}_\theta \frac{1}{\|X\|^2}.$$

证明：

首先计算 $\mathbb{E}_\theta (T_i^* - \theta_i)^2$：

$$\begin{aligned}
\mathbb{E}_\theta (T_i^* - \theta_i)^2 &= \mathbb{E}_\theta \left[\left(1 - \frac{b}{\|X\|^2}\right) X_i - \theta_i\right]^2 \\
&= \mathbb{E}_\theta \left[(X_i - \theta_i) - \frac{b}{\|X\|^2} X_i\right]^2 \\
&= \mathbb{E}_\theta \left[(X_i - \theta_i)^2 + b^2 \frac{X_i^2}{\|X\|^4} - 2b \frac{X_i (X_i - \theta_i)}{\|X\|^2}\right] \\
&= 1 + b^2 \mathbb{E}_\theta \frac{X_i^2}{\|X\|^4} - 2b \mathbb{E}_\theta \frac{X_i (X_i - \theta_i)}{\|X\|^2}.
\end{aligned}$$

分析最后一项的期望（以 $i = 1$ 为例）：

$$\begin{aligned}\mathbb{E}_\theta \frac{X_1\left(X_1-\theta_1\right)}{\|X\|^2} &= \int \frac{x_1\left(x_1-\theta_1\right)}{\|x\|^2} \prod_{i=1}^p \phi\left(x_i-\theta_i\right) dx_i \\ &= -\int \frac{x_1}{\|x\|^2} d\phi\left(x_1-\theta_1\right) \prod_{i=2}^p \phi\left(x_i-\theta_i\right) dx_i \\ &= \int \phi\left(x_1-\theta_1\right) d\left(\frac{x_1}{\|x\|^2}\right) \prod_{i=2}^p \phi\left(x_i-\theta_i\right) dx_i \\ &= \mathbb{E}_\theta\left[\frac{1}{\|X\|^2}-2 \frac{X_1^2}{\|X\|^4}\right].\end{aligned}$$

类似计算可以应用到所有 i.代入上述结果得到：

$$\mathbb{E}_\theta\left(T_i^*-\theta_i\right)^2=1+\left(b^2+4b\right) \mathbb{E}_\theta \frac{X_i^2}{\|X\|^4}-2 b \mathbb{E}_\theta \frac{1}{\|X\|^2}.$$

因此：

$$\begin{aligned}R\left(\theta, T^*\right) &= p+\left(b^2+4b\right) \mathbb{E}_\theta \frac{\sum_{i=1}^p X_i^2}{\|X\|^4}-2 b p \mathbb{E}_\theta \frac{1}{\|X\|^2} \\ &= p-\left[2 b(p-2)-b^2\right] \mathbb{E}_\theta \frac{1}{\|X\|^2}.\end{aligned}$$

结论：

Stein 估计量使用了其他独立观测值 $X_j(j \neq i)$，尽管这些观测值对 θ_i 独立，且它们的分布与 θ_i 无关.选择 $b=p-2$ 可以最大化对样本均值估计量 X 的改进，此时 Stein 估计量为：

$$T^*=\left[1-\frac{p-2}{\|X\|^2}\right] X.$$

备注：

1. Stein 估计量本身也是不可容许的.
2. 当 $\|\theta\| \rightarrow \infty$ 时，$R\left(\theta, T^*\right) \approx R(\theta, X)$.

第十二章

线性模型

假设有 n 个独立观测值 $Y_1, \ldots, Y_n$.这里不假定它们具有相同的分布.令 $X \in \mathbb{R}^{n \times p}$ 是一个给定的矩阵，其中的元素为 $\{x_{i, j}: i=1, \ldots, n, j=1, \ldots, p\}$.矩阵 X 被视为（固定的）输入，向量 $Y=\left(Y_1, \ldots, Y_n\right)^T$ 被视为（随机的）输出.矩阵 X 的列称为协变量（co-variables），并且矩阵 X 被称为设计矩阵（design matrix）.我们假设 X 是非随机的，即考虑固定设计的情况.

12.1 最小二乘估计的定义

我们的目标是根据 X 对 Y 进行预测，并决定使用线性逼近的方法：寻找 Y_i 关于 $x_{i,1},\ldots,x_{i,p}$ 的最佳线性逼近.通过残差平方和（residual sum of squares）来衡量拟合优度，这意味着我们需要最小化：

$$\sum_{i=1}^{n}\left(Y_i-a-\sum_{j=1}^{p}x_{i,j}b_j\right)^2$$

其中 $a\in\mathbb{R}$，$b=(b_1,\ldots,b_p)^T\in\mathbb{R}^p$.

为了简化表达，我们重新命名相关量.定义对所有 i，令 $x_{i,p+1}:=1$，并定义 $b_{p+1}:=a$.因此，对于任意 i，有 $a+\sum_{j=1}^{p}x_{i,j}b_j=\sum_{j=1}^{p+1}x_{i,j}b_j$.换句话说，如果在矩阵 X 中添加一列全为 1 的列，我们可以省略常数项 a.于是，将这一列全为 1 的列放在最前面，并将 $p+1$ 替换为 p，得到新的设计矩阵：

$$X:=\begin{pmatrix}1 & x_{1,2} & \cdots & x_{1,p}\\ 1 & x_{2,2} & \cdots & x_{2,p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n,2} & \cdots & x_{n,p}\end{pmatrix}$$

此时，我们最小化：

$$\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p}x_{i,j}b_j\right)^2$$

其中 $b=(b_1,\ldots,b_p)^T\in\mathbb{R}^p$.

令向量 $v\in\mathbb{R}^n$ 的欧几里得范数为：

$$\|v\|_2:=\sqrt{\sum_{i=1}^{n}v_i^2}$$

写作

$$Y=\begin{pmatrix}Y_1\\ \vdots\\ Y_n\end{pmatrix}$$

则

$$\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p}x_{i,j}b_j\right)^2=\|Y-Xb\|_2^2$$

定义 12.1.1

若 X 的秩为 p，称

$$\hat{\beta} := \arg\min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2$$

为最小二乘估计（least squares estimator）.我们称通过（线性）回归（regression）将 Y 对 X 进行拟合得到 $\hat{\beta}$.

Y 到列空间 $\{Xb : b \in \mathbb{R}^p\}$ 的距离通过将 Y 投影到该空间来最小化.实际上，有：

$$\frac{1}{2}\frac{\partial}{\partial b}\|Y - Xb\|_2^2 = -X^T(Y - Xb)$$

由此得到 $\hat{\beta}$ 满足所谓的正规方程（normal equations）：

$$X^T(Y - X\hat{\beta}) = 0$$

或者写为：

$$X^T Y = X^T X\hat{\beta}$$

若 X 的秩为 p，矩阵 $X^T X$ 存在逆矩阵 $\left(X^T X\right)^{-1}$，因此得到：

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

Y 在 $\{Xb : b \in \mathbb{R}^p\}$ 上的投影为：

$$\underbrace{X\left(X^T X\right)^{-1} X^T}_{\text{投影矩阵}} Y$$

回忆投影是形式为 PP^T 的线性映射，其中 $P^T P = I$.

我们可以写作 $X\left(X^T X\right)^{-1} X^T := PP^T$.

例子：$p = 1$ 的情况

若 $p = 1$，则

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

此时：

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix},$$

$$\left(X^T X\right)^{-1} = \left(n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2\right)^{-1} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

继续推导可得估计 $\hat{\alpha}$ 和 $\hat{\beta}$ 的具体表达式，如上所述.最后，我们可以通过模拟数据验证其性能，例如：

$$Y = 0.3 + 0.6x + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \frac{1}{4}\right), \quad \hat{\alpha} = 0.19, \hat{\beta} = 0.740$$

12.2 插曲：χ^2 分布

令 $Z_1, \ldots, Z_p$ 是独立同分布的 $\mathcal{N}(0,1)$ 随机变量，定义 p 维向量：

$$Z := \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix}.$$

则 Z 服从 $\mathcal{N}(0, I)$ 分布，其中 I 是 $p \times p$ 的单位矩阵.具有 p 自由度的 χ^2 分布定义为以下随机变量的分布：

$$\|Z\|_2^2 := \sum_{j=1}^{p} Z_j^2.$$

记号：$\|Z\|_2^2 \sim \chi_p^2$.

对于对称正定矩阵 Σ，可以定义其平方根 $\Sigma^{1/2}$ 为一个对称正定矩阵，满足：

$$\Sigma^{1/2}\Sigma^{1/2} = \Sigma.$$

其逆记为 $\Sigma^{-1/2}$（是 Σ^{-1} 的平方根）.如果 $Z \in \mathbb{R}^p$ 服从 $\mathcal{N}(0, \Sigma)$ 分布，则变换后的向量：

$$\tilde{Z} := \Sigma^{-1/2} Z$$

服从 $\mathcal{N}(0, I)$ 分布.因此有：

$$Z^T \Sigma^{-1} Z = \tilde{Z}^T \tilde{Z} = \|\tilde{Z}\|_2^2 \sim \chi_p^2.$$

12.3 最小二乘估计的分布

定义 12.3.1

对于 $f = EY$，令 $\beta^* := (X^T X)^{-1} X^T f$，称 $X\beta^*$ 为 f 的最佳线性逼近（best linear approximation）.

引理 12.3.1

假设 $E\epsilon\epsilon^T = \sigma^2 I$，其中 $\epsilon := Y - f$.则：

1. $E\hat{\beta} = \beta^*$，$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$；
2. $E\|X(\hat{\beta} - \beta^*)\|_2^2 = \sigma^2 p$；
3. $E\|X\hat{\beta} - f\|_2^2 = \underbrace{\sigma^2 p}_{\text{估计误差（estimation error）}} + \underbrace{\|X\beta^* - f\|_2^2}_{\text{模型误差（misspecification error）}}$.

证明

1. **期望和协方差**

   由直接计算可得：

$$\hat{\beta} - \beta^* = \underbrace{(X^T X)^{-1} X^T}_{:=A} \epsilon.$$

   因此：

$$E(\hat{\beta} - \beta^*) = AE\epsilon = 0,$$

   并且 $\hat{\beta}$ 的协方差矩阵为：

$$\begin{aligned} \operatorname{Cov}(\hat{\beta}) &= \operatorname{Cov}(A\epsilon) \\ &= A \underbrace{\operatorname{Cov}(\epsilon)}_{=\sigma^2 I} A^T \\ &= \sigma^2 A A^T = \sigma^2 (X^T X)^{-1}. \end{aligned}$$

2. **估计误差**

   定义投影矩阵 $PP^T := X(X^T X)^{-1} X^T$，则：

$$\|X(\hat{\beta} - \beta^*)\|_2^2 = \|PP^T \epsilon\|_2^2 := \sum_{j=1}^{p} V_j^2,$$

   其中 $V := P^T \epsilon$，并且：

$$EV = P^T E\epsilon = 0,$$

   且协方差：

$$\operatorname{Cov}(V) = P^T \operatorname{Cov}(\epsilon) P = \sigma^2 I.$$

   因此：

$$E \sum_{j=1}^{p} V_j^2 = \sum_{j=1}^{p} E V_j^2 = \sigma^2 p.$$

3. **总误差**

   对于任意 b，由勾股定理可得：

$$\|Xb - f\|_2^2 = \|X(b - \beta^*)\|_2^2 + \|X\beta^* - f\|_2^2,$$

   因为 $X\beta^* - f$ 与 X 正交.

   故：

$$E\|X\hat{\beta} - f\|_2^2 = E\|X(\hat{\beta} - \beta^*)\|_2^2 + \|X\beta^* - f\|_2^2 = \sigma^2 p + \|X\beta^* - f\|_2^2.$$

引理 12.3.2

假设 $\epsilon := Y - f \sim \mathcal{N}(0, \sigma^2 I)$，则有：

1. $\hat{\beta} - \beta^* \sim \mathcal{N}\left(0, \sigma^2 \left(X^T X\right)^{-1}\right)$；
2. $\frac{\left\|X\left(\hat{\beta}-\beta^*\right)\right\|_2^2}{\sigma^2} \sim \chi_p^2$，其中 χ_p^2 是具有 p 自由度的 χ^2 分布（定义见 12.2 节）.

证明

1. **关于 $\hat{\beta}$ 的分布**
   因为 $\hat{\beta}$ 是多元正态分布 ϵ 的线性函数，因此最小二乘估计 $\hat{\beta}$ 也服从多元正态分布.
2. **关于 χ^2 分布**
   定义投影矩阵 $PP^T := X\left(X^T X\right)^{-1} X^T$，则有：

$$\left\|X\left(\hat{\beta} - \beta^*\right)\right\|_2^2 = \left\|PP^T \epsilon\right\|_2^2 := \sum_{j=1}^{p} V_j^2$$

   其中 $V := P^T \epsilon$ 的分量是独立同分布的 $\mathcal{N}(0, \sigma^2)$ 随机变量.

备注

- **模型误差** $\|X\beta^* - f\|_2^2$ 源于可能的线性模型设定错误，即 f 可能不是 X 的列线性组合.这通常也被称为**近似误差**.
- **估计误差**是方差项 $\sigma^2 p$.

更一般地，许多估计量近似服从正态分布（例如样本中位数），而许多检验统计量的零分布近似为 χ^2 分布（例如 χ^2 拟合优度检验统计量）.这一现象的原因在于许多模型可以在某种意义上被线性模型近似，而许多负对数似然函数类似于最小二乘损失函数（参见第 14 章）.理解线性模型是理解更复杂模型的重要第一步.

推论 12.3.1

假设线性模型设定正确，即存在某个 $\beta \in \mathbb{R}^p$ 使得：

$$EY = X\beta$$

并假设 $\epsilon := Y - EY \sim \mathcal{N}(0, \sigma^2)$，其中 $\sigma^2 := \sigma_0^2$ 是已知的.那么关于 $H_0 : \beta = \beta_0$ 的检验为：

当

$$\frac{\left\|X\left(\hat{\beta} - \beta^0\right)\right\|_2^2}{\sigma_0^2} > G_p^{-1}(1 - \alpha)$$

时拒绝 H_0，其中 G_p 是具有 p 自由度的 χ_p^2 分布的分布函数.

备注

当 σ^2 未知时，可以使用以下估计量估计 σ^2：

$$\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|_2^2}{n-p},$$

其中 $\hat{\epsilon} := Y - X\hat{\beta}$ 是残差向量.在前述假设下（但 σ^2 可能未知），检验统计量：

$$\frac{\|X(\hat{\beta} - \beta^0)\|_2^2}{p\hat{\sigma}^2}$$

服从具有 p 和 $n-p$ 自由度的 F 分布.

12.4 插曲：一些矩阵代数

令 $z \in \mathbb{R}^p$ 为向量，$B \in \mathbb{R}^{q\times p}$ 为 $q \times p$ 矩阵（$p \geq q$），且矩阵 B 的秩为 q.令 $V \in \mathbb{R}^{p\times p}$ 是正定矩阵.

引理 12.4.1

有以下等式：

$$\max_{a\in\mathbb{R}^p : Ba=0} \left\{2a^T z - a^T a\right\} = z^T z - z^T B^T \left(BB^T\right)^{-1} Bz$$

证明

通过引入拉格朗日乘子 $\lambda \in \mathbb{R}^p$，我们有：

$$\frac{\partial}{\partial a}\left\{2a^T z - a^T a + 2a^T B^T \lambda\right\} = z - a + B^T\lambda$$

令

$$a_* := \arg\max_{a\in\mathbb{R}^p; Ba=0} \left\{2a^T z - a^T a\right\},$$

则有：

$$z - a_* + B^T\lambda = 0,$$

即：

$$a_* = z + B^T\lambda.$$

结合约束条件 $Ba_* = 0$ 得到：

$$Bz + BB^T\lambda = 0 \implies \lambda = -\left(BB^T\right)^{-1}Bz.$$

代入 a_*，可得：

$$a_* = z - B^T\left(BB^T\right)^{-1}Bz.$$

因此：

$$\begin{aligned} a_*^T a_* &= \left(z^T - z^T B^T\left(BB^T\right)^{-1}B\right)\left(z - B^T\left(BB^T\right)^{-1}Bz\right) \\ &= z^T z - z^T B^T\left(BB^T\right)^{-1}Bz, \end{aligned}$$

从而得到：

$$2a_*^T z - a_*^T a_* = z^T z - z^T B^T\left(BB^T\right)^{-1}Bz.$$

引理 12.4.2

有以下等式：

$$\max_{a\in\mathbb{R}^p : Ba=0}\left\{2a^T z - a^T V a\right\} = z^T V^{-1} z - z^T V^{-1} B^T\left(BV^{-1}B^T\right)^{-1}BV^{-1}z.$$

证明

令 $b := V^{1/2}a$，$y := V^{-1/2}z$，$C := BV^{-1/2}$，则：

$$\max_{a:Ba=0}\left\{2a^T z - a^T V a\right\} = \max_{b:Cb=0}\left\{2b^T y - b^T b\right\}.$$

由引理 12.4.1，可得：

$$\begin{aligned} &\max_{b:Cb=0}\left\{2b^T y - b^T b\right\} \\ &= y^T y - y^T C^T\left(CC^T\right)^{-1}Cy \\ &= z^T V^{-1} z - z^T V^{-1} B^T\left(BV^{-1}B^T\right)^{-1}BV^{-1}z. \end{aligned}$$

推论 12.4.1

令 $L(a) := 2a^T z - a^T V a$.则 $L(a)$ 在无约束最大值与约束最大值之间的差为：

$$\max_a L(a) - \max_{a:Ba=0} L(a) = z^T V^{-1} B^T\left(BV^{-1}B^T\right)^{-1}BV^{-1}z.$$

12.5 检验线性假设

在本节中，我们假设模型如下：

$$Y = X\beta + \epsilon,$$

其中 $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

我们希望检验以下假设：

$$H_0 : B\beta = 0,$$

其中 $B \in \mathbb{R}^{q\times p}$ 是给定的 $q \times p$ 矩阵.

定义

$$\hat{\beta}_0 := \arg \min_{b\in\mathbb{R}^p : Bb=0} \|Y - Xb\|_2^2$$

为在约束条件 $Bb = 0$ 下的最小二乘估计量.

引理 12.5.1

在 H_0 成立的条件下，

$$\frac{\|Y - X\hat{\beta}_0\|_2^2 - \|Y - X\hat{\beta}\|_2^2}{\sigma^2}$$

服从 χ_q^2 分布.

证明

由于

$$\|Y - Xb\|_2^2 = \|\epsilon\|_2^2 - 2\epsilon^T X(b - \beta) + (b - \beta)^T X^T X(b - \beta),$$

在 H_0 成立的情况下（令 $\tilde{b} := b - \beta$），有

$$\hat{\beta}_0 - \beta = \arg \max_{\tilde{b}\in\mathbb{R}^p : B\tilde{b}=0} \left\{2\epsilon^T X\tilde{b} - \tilde{b}^T X^T X\tilde{b}\right\}.$$

利用推论 12.4.1 可得：

$$\begin{aligned}\|Y - X\hat{\beta}_0\|_2^2 - \|Y - X\hat{\beta}\|_2^2 &= \underbrace{\epsilon^T X(X^T X)^{-1}B^T}_{:=Z^T}\left(B(X^T X)^{-1}B^T\right)^{-1}\underbrace{B(X^T X)^{-1}X^T\epsilon}_{:=Z} \\ &= Z^T\left(B(X^T X)^{-1}B^T\right)^{-1}Z.\end{aligned}$$

向量

$$Z := B(X^T X)^{-1}X^T\epsilon$$

是一个 q 维的多元正态分布，均值为 0，协方差矩阵为：

$$\sigma^2\left(B(X^T X)^{-1}X^T\right)\left(B(X^T X)^{-1}X^T\right)^T = \sigma^2 B(X^T X)^{-1}B^T.$$

因此，在 H_0 成立时，

$$\frac{\|Y - X\hat{\beta}_0\|_2^2 - \|Y - X\hat{\beta}\|_2^2}{\sigma^2}$$

服从 χ_q^2 分布.

第 13 章

渐近理论

在本章及之后的章节中，观测值 $X_1, \ldots, X_n$ 被看作是无限序列 $X_1, \ldots, X_n, \ldots$ 中的前 n 个样本，这些随机变量独立同分布，取值于 $\mathcal{X}$，且分布为 P.我们称 X_i 是某随机变量 $X \in \mathcal{X}$ 的 i.i.d. 样本，其分布为 P.令 $\mathbb{P} = P \times P \times \cdots$ 为整个序列 $\{X_i\}_{i=1}^{\infty}$ 的分布.

参数模型为：

$$\mathcal{P} := \{P_\theta : \theta \in \Theta\}.$$

当 $P = P_\theta$ 时，记 $\mathbb{P} = \mathbb{P}_\theta = P_\theta \times P_\theta \times \cdots$.感兴趣的参数为：

$$\gamma := g(\theta) \in \mathbb{R}^p,$$

其中 $g : \Theta \to \mathbb{R}^p$ 是给定函数.记

$$\Gamma := \{g(\theta) : \theta \in \Theta\}$$

为 γ 的参数空间.

估计量

令

$$T_n(X_1, \ldots, X_n)$$

为基于样本 $X_1, \ldots, X_n$ 的某估计量，其为样本的函数.简记为：

$$T_n := T_n(X_1, \ldots, X_n).$$

假设估计量 T_n 对所有 n 都定义，即实际上考虑的是估计量序列 $\{T_n\}_{n=1}^{\infty}$.我们感兴趣的是参数 γ 的估计量 $T_n \in \Gamma$.

备注

在 i.i.d. 假设下，自然假设每个 T_n 是样本的对称函数，即：

$$T_n(X_1, \ldots, X_n) = T_n(X_{\pi_1}, \ldots, X_{\pi_n}),$$

对所有 $\{1, \ldots, n\}$ 的排列 π 成立.在这种情况下，可以将 T_n 写成 $T_n = Q(\hat{P}_n)$，其中 $\hat{P}_n$ 是经验分布（详见 2.4.1 节）.

13.1 收敛类型

定义 13.1.1

令 $\{Z_n\}_{n=1}^{\infty}$ 和 Z 是定义在同一概率空间上的 $\mathbb{R}^p$ 值随机变量.当对于任意 $\epsilon > 0$，有

$$\lim_{n\to\infty} \mathbb{P}(\|Z_n - Z\| > \epsilon) = 0$$

时，称 Z_n 在概率意义下收敛到 Z.

记号：$Z_n \xrightarrow{\mathbb{P}} Z$.

备注

切比雪夫不等式可用于证明概率收敛.其表述为，对于所有递增函数 $\psi : [0,\infty) \to [0,\infty)$，有

$$\mathbb{P}(\|Z_n - Z\| \geq \epsilon) \leq \frac{\mathbb{E}\psi(\|Z_n - Z\|)}{\psi(\epsilon)}$$

定义 13.1.2

令 $\{Z_n\}_{n=1}^{\infty}$ 和 Z 是 $\mathbb{R}^p$ 值随机变量.如果对任意连续且有界的函数 f，

$$\lim_{n\to\infty} \mathbb{E}f(Z_n) = \mathbb{E}f(Z),$$

则称 Z_n 按分布收敛到 Z.

记号：$Z_n \xrightarrow{\mathcal{D}} Z$.

备注

概率收敛蕴含分布收敛，但反之不一定成立.

例 13.1.1 中心极限定理 (CLT)

令 $X_1, X_2, \ldots$ 是独立同分布的实值随机变量，均值为 μ，方差为 σ^2.令 $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ 为前 n 个样本的平均值.则根据中心极限定理 (CLT)，

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

即：

$$\mathbb{P}\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq z\right) \to \Phi(z), \forall z,$$

其中 $\Phi(z)$ 是标准正态分布函数.

定理 13.1.1 (Cramér-Wold 装置)

设 $\{Z_n\}, Z$ 是 $\mathbb{R}^p$ 值随机变量.则有:

$$Z_n \xrightarrow{\mathcal{D}} Z \Leftrightarrow a^T Z_n \xrightarrow{\mathcal{D}} a^T Z \quad \forall a \in \mathbb{R}^p.$$

例 13.1.2 多元中心极限定理 (Multivariate CLT)

令 $X_1, X_2, \ldots$ 是随机变量 $X = (X^{(1)}, \ldots, X^{(p)})^T$ 的独立同分布拷贝，$X \in \mathbb{R}^p$.假设 $EX = \mu = (\mu_1, \ldots, \mu_p)^T$ 且协方差矩阵 $\Sigma = \mathrm{Cov}(X) := EXX^T - \mu\mu^T$ 存在.则对于任意 $a \in \mathbb{R}^p$,

$$E(a^T X) = a^T \mu, \quad \mathrm{var}(a^T X) = a^T \Sigma a$$

定义样本均值向量

$$\bar{X}_n = (\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(p)})^T,$$

根据一维中心极限定理，对于任意 $a \in \mathbb{R}^p$，

$$\sqrt{n}\left(a^T \bar{X}_n - a^T \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, a^T \Sigma a).$$

利用 Cramér-Wold 装置，可得 p 维中心极限定理:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma).$$

定理 13.1.2 (Portmanteau 定理)

设 $\{Z_n\}, Z$ 是 $\mathbb{R}^p$ 值随机变量，Q 是 Z 的分布，$G = Q(Z \leq \cdot)$ 是其分布函数.以下陈述等价:

1. $Z_n \xrightarrow{\mathcal{D}} Z$，即 $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z)$ 对所有有界连续函数 f 成立.
2. $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z)$ 对所有有界 Lipschitz 函数 f 成立.
3. $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z)$ 对所有 Q -几乎处处连续的有界函数 f 成立.
4. $\mathbb{P}(Z_n \leq z) \to G(z)$ 对 G 连续点 z 成立.

13.1.1 随机阶符号

令 $\{Z_n\}$ 是一组 $\mathbb{R}^p$ 值随机变量，$\{r_n\}$ 是严格正的随机变量.当

$$\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{P}(\|Z_n\| > M) = 0$$

时，记作

$$Z_n = \mathcal{O}_{\mathbf{P}}(1),$$

即 Z_n 在概率意义下有界.这也称为序列 $\{Z_n\}$ 的一致紧性.如果 $Z_n/r_n = \mathcal{O}_{\mathbf{P}}(1)$，记作 $Z_n = \mathcal{O}_{\mathbf{P}}(r_n)$.

当 Z_n 在概率意义下收敛到 0 时，记作

$$Z_n = o_{\mathbf{P}}(1),$$

此外，若 $Z_n/r_n = o_{\mathbf{P}}(1)$，记作 $Z_n = o_{\mathbf{P}}(r_n)$，即 Z_n 在概率意义下是 r_n 的小阶.

13.1.2 收敛的一些推论

引理 13.1.1

若 Z_n 按分布收敛，则 $Z_n = \mathcal{O}_{\mathbf{P}}(1)$.

证明

为简化，取 $p = 1$（Cramér-Wold 装置）.设 $Z_n \xrightarrow{\mathcal{D}} Z$，其中 Z 的分布函数为 G.对于任意 G 的连续点 M，

$$\mathbb{P}(Z_n > M) \to 1 - G(M),$$

对于任意 G 的连续点 $-M$，

$$\mathbb{P}(Z_n \leq -M) \to G(-M).$$

由于 $1 - G(M)$ 和 $G(-M)$ 在 $M \to \infty$ 时均收敛到 0，因此结论成立.

例 13.1.3

平均值与其期望的差在概率意义下是 $1/\sqrt{n}$ 阶.
令 $X_1, X_2, \ldots$ 是随机变量 $X \in \mathbb{R}$ 的独立同分布拷贝，$EX = \mu$ 且 $\operatorname{var}(X) < \infty$.则根据中心极限定理，

$$\bar{X}_n - \mu = \mathcal{O}_{\mathbf{P}}(1/\sqrt{n}).$$

定理 13.1.3 (Slutsky 定理)

设 $\{Z_n, A_n\}, Z$ 是一组 $\mathbb{R}^p$ 值随机变量，$a \in \mathbb{R}^p$ 是常数向量.若 $Z_n \xrightarrow{\mathcal{D}} Z, A_n \xrightarrow{\mathbb{P}} a$，则

$$A_n^T Z_n \xrightarrow{\mathcal{D}} a^T Z.$$

证明

取一个有界 Lipschitz 函数 f，满足

$$|f| \leq C_B, \quad |f(z) - f(\tilde{z})| \leq C_L \|z - \tilde{z}\|.$$

则

$$\begin{aligned}\left|\mathbb{E}f(A_n^TZ_n)-\mathbb{E}f(a^TZ)\right| \le &\left|\mathbb{E}f(A_n^TZ_n)-\mathbb{E}f(a^TZ_n)\right| \\ &+\left|\mathbb{E}f(a^TZ_n)-\mathbb{E}f(a^TZ)\right|.\end{aligned}$$

由于函数 $z\mapsto f(a^Tz)$ 是有界 Lipschitz 的，其 Lipschitz 常数为 $\|a\|C_L$，第二项收敛到 0.对于第一项，定义 $S_n=\{\|Z_n\|\le M,\|A_n-a\|\le\epsilon\}$，则

$$\left|\mathbb{E}f(A_n^TZ_n)-\mathbb{E}f(a^TZ_n)\right|\le C_L\epsilon M+2C_B\mathbb{P}(S_n^c).$$

利用 $\mathbb{P}(S_n^c)$ 可通过适当选择 ϵ 小、M 大和 n 大而变得任意小.

13.2 一致性与渐近正态性

定义 13.2.1
若一组估计量 $\{T_n\}$ 满足

$$T_n\xrightarrow{\mathbb{P}_\rho}\gamma,$$

称其是一致估计量.

定义 13.2.2
若一组估计量 $\{T_n\}$ 满足

$$\sqrt{n}(T_n-\gamma)\xrightarrow{\mathcal{D}_\theta}\mathcal{N}(0,V_\theta),$$

则称其具有渐近正态性，渐近协方差矩阵为 V_θ.

例 13.2.1 平均值的一致性与渐近正态性
设 $\mathcal{P}$ 为位置模型：

$$\mathcal{P}=\{P_{\mu,F_0}(X\le\cdot):=F_0(\cdot-\mu),\mu\in\mathbb{R},F_0\in\mathcal{F}_0\}.$$

参数为 $\theta=(\mu,F_0)$，参数空间为 $\Theta=\mathbb{R}\times\mathcal{F}_0$.假设对于所有 $F_0\in\mathcal{F}_0$，

$$\int xdF_0(x)=0,\quad \sigma_{F_0}^2:=\int x^2dF_0(x)<\infty.$$

令 $g(\theta)=\mu$，$T_n:=\frac{1}{n}\sum_{i=1}^n X_i=\bar{X}_n$.根据大数定律，$T_n$ 是 μ 的一致估计量；根据中心极限定理，

$$\sqrt{n}(\bar{X}_n-\mu)\xrightarrow{\mathcal{D}_\theta}\mathcal{N}(0,\sigma_{F_0}^2).$$

13.3 渐近线性性

对于许多估计量，渐近正态性是渐近线性性的结果，即估计量可以近似为一个平均值，从而可以应用中心极限定理.

定义 13.3.1

估计量序列 $\{T_n\}$ 对 $\gamma = g(\theta) \in \mathbb{R}^p$ 被称为渐近线性，如果存在函数 $l_\theta : \mathcal{X} \to \mathbb{R}^p$，使得 $\mathbb{E}_\theta l_\theta(X) = 0$ 且

$$\mathbb{E}_\theta l_\theta(X) l_\theta^T(X) =: V_\theta < \infty,$$

并满足

$$T_n - \gamma = \frac{1}{n} \sum_{i=1}^{n} l_\theta(X_i) + o_{\mathbf{P}_\theta}(1/\sqrt{n}).$$

备注

函数 l_θ 称为 T_n 的影响函数 (influence function).它近似衡量了额外观测 x 对估计量的影响.

例 13.3.1 样本均值的影响函数

若 X 的分量具有有限方差，估计量 $T_n := \bar{X}_n$ 是均值 μ 的线性且渐近线性估计量，其影响函数为:

$$l_\theta(x) = x - \mu.$$

例 13.3.2 样本方差的影响函数

令 X 是实值随机变量，满足 $\mathbb{E}_\theta X =: \mu, \mathrm{var}_\theta(X) =: \sigma^2$，且四阶矩 $\kappa := \mathbb{E}_\theta (X - \mu)^4$ 存在.样本方差为:

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n \right)^2.$$

估计量 $\hat{\sigma}_n^2$ 为:

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}_n \right)^2.$$

可以验证:

$$S^2 - \hat{\sigma}^2 = \mathcal{O}_{\mathbf{P}}(1/n) = o_{\mathbf{P}}(1/\sqrt{n}).$$

重写 $\hat{\sigma}_n^2$:

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 - \left(\bar{X}_n - \mu \right)^2.$$

由中心极限定理，$\bar{X}_n - \mu = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$，因此:

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + \mathcal{O}_{\mathbf{P}_\theta}(1/n).$$

由此可知，$\hat{\sigma}_n^2$（以及 S^2）是渐近线性的，其影响函数为：

$$l_\theta(x) = (x-\mu)^2 - \sigma^2.$$

渐近方差为：

$$V_\theta = \mathbb{E}_\theta\big((X-\mu)^2-\sigma^2\big)^2 = \kappa - \sigma^4.$$

13.4 δ -技术

定理 13.4.1

令 $\{T_n\}, Z$ 是 $\mathbb{R}^p$ 上的随机变量，$c \in \mathbb{R}^p$ 是非随机向量，$\{r_n\}$ 是严格正的非随机数列，且 $r_n \downarrow 0$.设 $h : \mathbb{R}^p \to \mathbb{R}$ 在 c 可导，其导数为 $\dot{h}(c) \in \mathbb{R}^p$.若

$$\frac{T_n - c}{r_n} \xrightarrow{\mathcal{D}} Z,$$

则

$$\frac{h(T_n)-h(c)}{r_n} \xrightarrow{\mathcal{D}} \dot{h}(c)^T Z,$$

并且

$$h(T_n) - h(c) = \dot{h}(c)^T(T_n - c) + o_{\mathbf{P}}(r_n).$$

推论 13.4.1

若 T_n 是 $\gamma = g(\theta) \in \mathbb{R}^p$ 的渐近正态估计量，其渐近协方差矩阵为 V_θ，且 h 在 γ 可导，则 $h(T_n)$ 是 $h(\gamma)$ 的渐近正态估计量，其渐近方差为：

$$\dot{h}(\gamma)^T V_\theta \dot{h}(\gamma).$$

如果 T_n 是 γ 的渐近线性估计量，其影响函数为 l_θ，则 $h(T_n)$ 是 $h(\gamma)$ 的渐近线性估计量，其影响函数为：

$$\dot{h}(\gamma)^T l_\theta.$$

例 13.4.1 指数分布参数的渐近线性估计量

令 $X_1, \ldots, X_n$ 是来自 Exponential(θ) 分布的样本，其中 $\theta > 0$.

$\bar{X}_n$ 是 $E_\theta X = 1/\theta = \gamma$ 的线性估计量，其影响函数为 $l_\theta(x) = x - 1/\theta$.$\sqrt{n}(T_n - 1/\theta)$ 的方差为 $1/\theta^2 = \gamma^2$.根据定理 13.4.1，$1/\bar{X}_n$ 是 θ 的渐近线性估计量.此时，$h(\gamma) = 1/\gamma$，因此

$\dot{h}(\gamma) = -1/\gamma^2.1/\bar{X}_n$ 的影响函数为：

$$\dot{h}(\gamma) l_\theta(x) = -\frac{1}{\gamma^2}(x - \gamma) = -\theta^2(x - 1/\theta)$$

$1/\bar{X}_n$ 的渐近方差为：

$$[\dot{h}(\gamma)]^2 \gamma^2 = \frac{1}{\gamma^2} = \theta^2$$

因此：

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \theta\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, \theta^2\right).$$

例 13.4.2 样本均值和样本方差的二维渐近线性性

回到例 13.3.2，令 X 为实值随机变量，满足 $E_\theta X = \mu, \mathrm{var}_\theta(X) = \sigma^2$，且 $\kappa := E_\theta(X - \mu)^4$ 存在.定义第 r 阶矩 $\mu_r := E_\theta X^r$ $(r = 1, 2, 3, 4)$.考虑估计量：

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

我们有：

$$\hat{\sigma}_n^2 = h(T_n)$$

其中 $T_n = \begin{pmatrix} T_{n,1} \\ T_{n,2} \end{pmatrix}$，且：

$$T_{n,1} = \bar{X}_n, \quad T_{n,2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

函数 h 定义为：

$$h(t) = t_2 - t_1^2, \quad t = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}$$

估计量 T_n 的影响函数为：

$$l_\theta(x) = \begin{pmatrix} x - \mu_1 \\ x^2 - \mu_2 \end{pmatrix}$$

根据二维中心极限定理：

$$\sqrt{n}\left(T_n - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, V_\theta\right),$$

其中：

$$V_\theta = \begin{pmatrix} \mu_2 - \mu_1^2 & \mu_3 - \mu_1\mu_2 \\ \mu_3 - \mu_1\mu_2 & \mu_4 - \mu_2^2 \end{pmatrix}$$

我们有：

$$\dot{h}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right) = \begin{pmatrix}-2\mu_1\\ 1\end{pmatrix}$$

因此，$\hat{\sigma}_n^2$ 的影响函数为：

$$\dot{h}^T\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right) l_\theta(x) = \begin{pmatrix}-2\mu_1\\ 1\end{pmatrix}^T \begin{pmatrix} x-\mu_1 \\ x^2-\mu_2 \end{pmatrix} = (x-\mu)^2 - \sigma^2$$

计算得：

$$\begin{pmatrix}-2\mu_1\\ 1\end{pmatrix}^T V_\theta \begin{pmatrix}-2\mu_1\\ 1\end{pmatrix} = \kappa - \sigma^4$$

即 δ -方法与例 13.3.2 的直接方法得到相同结果.

第 14 章

M-估计量

回忆第 2.4.3 节定义的最大似然估计量 (MLE).在本章中，我们引入一类广义的估计量，MLE 是其中的一个特例.它们是某些经验风险函数的最小化解.

令每个 $\gamma \in \Gamma$ 关联一个损失函数 $\rho_\gamma(X)$，例如在第 10 章中构造的损失函数：假设 $L(\theta, a)$ 表示选择动作 a 时的损失.我们固定某个决策函数 $d(x)$，并重写为：

$$L(\theta, d(x)) := \rho_\gamma(x)$$

假设损失 L 仅通过感兴趣的参数 $\gamma = g(\theta)$ 依赖于 θ.我们要求理论风险：

$$\mathcal{R}(c) := E_\theta \rho_c(X)$$

在 $c = \gamma$ 处最小化，即：

$$\gamma = \arg\min_{c\in\Gamma} E_\theta \rho_c(X) = \arg\min_{c\in\Gamma} \mathcal{R}(c).$$

如果 $c \mapsto \rho_c(x)$ 对所有 x 可微，则定义：

$$\psi_c(x) := \dot{\rho}_c(x) := \frac{\partial}{\partial c}\rho_c(x).$$

假设期望与求导可以交换，则有：

$$\dot{\mathcal{R}}(\gamma) = 0,$$

其中 $\dot{\mathcal{R}}(c) = E_\theta \psi_c(X)$.定义经验风险为：

$$\hat{\mathcal{R}}_n(c) := \frac{1}{n}\sum_{i=1}^{n} \rho_c(X_i), \quad c \in \Gamma.$$

[3]### 定义 14.0.1 M 估计量

M 估计量 $\hat{\gamma}_n$ 定义为：

$$\hat{\gamma}_n := \arg\min_{c\in\Gamma} \frac{1}{n}\sum_{i=1}^{n} \rho_c(X_i) = \arg\min_{c\in\Gamma} \hat{\mathcal{R}}_n(c).$$

"M 估计量"中的“M”代表最小化（Minimizer，或若考虑负号则为最大化）.

如果 $\rho_c(x)$ 关于 c 对所有 x 可微，则通常可以通过设导数为零来定义 $\hat{\gamma}_n$：

$$\dot{\hat{\mathcal{R}}}_n(c) = \frac{\partial}{\partial c}\frac{1}{n}\sum_{i=1}^{n} \rho_c(X_i) = \frac{1}{n}\sum_{i=1}^{n} \psi_c(X_i)$$

该方法称为 Z 估计量."Z"代表零（Zero）.

定义 14.0.2 Z 估计量

Z 估计量 $\hat{\gamma}_n$ 定义为下列方程的解：

$$\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0,$$

其中 $\dot{\hat{\mathcal{R}}}_n(c) = \frac{1}{n}\sum_{i=1}^{n} \psi_c(X_i)$.
注：假设解 $\hat{\gamma}_n \in \Gamma$ 存在.

例 14.0.1 最小二乘估计量

令 $X \in \mathbb{R}$，感兴趣的参数为 $\mathbb{E}_\theta X$ 的均值 μ.假设 X 具有有限方差 σ^2.则：

$$\mu = \arg\min_c \mathbb{E}_\theta (X-c)^2,$$

因为根据偏差-方差分解公式：

$$\mathbb{E}_\theta (X-c)^2 = \sigma^2 + (\mu - c)^2.$$

因此此时可以令：

$$\rho_c(x) = (x-c)^2.$$

显然：

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-c)^2$$

在 $c=\bar{X}_n:=\sum_{i=1}^{n}X_i/n$ 处达到最小值.参见第 2.3 节.

14.1 MLE 作为 M 估计量的特例

假设 $\Theta\subset\mathbb{R}^p$ 且密度 $p_\theta=dP_\theta/d\nu$ 关于某 σ -有限测度 ν 存在.

定义 14.1.1 相对熵（Kullback-Leibler 信息）

定义：

$$K(\tilde{\theta}\mid\theta)=\mathbb{E}_\theta\log\left(\frac{p_\theta(X)}{p_{\tilde{\theta}}(X)}\right),$$

称为 Kullback-Leibler 信息或相对熵.

注：需要注意避免分母为零的情况！例如，可以假设支持集 $\{x:p_\theta(x)>0\}$ 不依赖于 θ（见第 5 章 CRLB 的条件 I）.

取对所有 $\tilde{\theta}\in\Theta$：

$$\rho_{\tilde{\theta}}(x)=-\log p_{\tilde{\theta}}(x).$$

于是：

$$\mathcal{R}(\tilde{\theta})=-\mathbb{E}_\theta\log p_{\tilde{\theta}}(X).$$

容易验证：

$$K(\tilde{\theta}\mid\theta)=\mathcal{R}(\tilde{\theta})-\mathcal{R}(\theta).$$

引理 14.1.1

函数 $\mathcal{R}(\tilde{\theta})=-\mathbb{E}_\theta\log p_{\tilde{\theta}}(X)$ 在 $\tilde{\theta}=\theta$ 处取最小值：

$$\theta=\arg\min_{\tilde{\theta}}\mathcal{R}(\tilde{\theta}).$$

证明：

证明 $K(\tilde{\theta}\mid\theta)\geq 0$ 即可.由 Jensen 不等式（对数函数是凹函数）可得：

$$
\begin{aligned}
K(\tilde{\theta} \mid \theta) &= -\mathbb{E}_\theta \log\left(\frac{p_{\tilde{\theta}}(X)}{p_\theta(X)}\right) \\
&\geq -\log\left(\mathbb{E}_\theta\left(\frac{p_{\tilde{\theta}}(X)}{p_\theta(X)}\right)\right) \\
&= -\log 1 = 0.
\end{aligned}
$$

当 $\rho_{\tilde{\theta}}(x) = -\log p_{\tilde{\theta}}(x)$ 时，$\psi_{\tilde{\theta}}(x) := \dot{\rho}_{\tilde{\theta}}(x) = -s_{\tilde{\theta}}(x)$，其中 s_θ 为得分函数：

$$
s_\theta = \dot{p}_\theta / p_\theta,
$$

参见定义 4.7.1.根据引理 4.7.1，$\mathbb{E}_\theta s_\theta(X) = 0$.这表明 θ 是方程：

$$
\dot{\mathcal{R}}(\theta) = 0,
$$

的解，其中 $\dot{\mathcal{R}}(\tilde{\theta}) = \mathbb{E}_\theta \psi_{\tilde{\theta}}(X)$，且 $\psi_{\tilde{\theta}} = -\dot{p}_{\tilde{\theta}} / p_{\tilde{\theta}}$.

当 $\rho_{\tilde{\theta}}(x) = -\log p_{\tilde{\theta}}(x)$ 时，M 估计量就是最大似然估计量（MLE）：

$$
\hat{\theta} = \arg\min_{\tilde{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left(-\log p_{\tilde{\theta}}(X_i)\right).
$$

14.2 M 估计量的一致性

需要注意的是，γ 最小化了一个理论期望，而 M 估计量 $\hat{\gamma}_n$ 最小化了经验平均.同样地，γ 是一个使理论期望为零的解，而 Z 估计量 $\hat{\gamma}_n$ 是使经验平均为零的解.

根据大数定律，平均值收敛到期望值.因此，M 估计量（或 Z 估计量）是有意义的.然而，一致性和进一步的性质并不是立即显然的，因为我们实际上需要对一系列参数值 $c \in \Gamma$ 同时实现从平均值到期望值的收敛.这涉及到经验过程理论.

我们将借用经验过程理论中的符号.对于一个函数 $f : \mathcal{X} \to \mathbb{R}$，定义：

$$
P_\theta f := \mathbb{E}_\theta f(X), \quad \hat{P}_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i).
$$

根据大数定律，如果 $P_\theta |f| < \infty$，则有：

$$
\left| \left(\hat{P}_n - P_\theta\right) f \right| \to 0, \quad \mathbb{P}_\theta - \text{几乎处处（a.s.）}.
$$

使用上述记号，我们可以表示：

$$
\hat{\mathcal{R}}_n(c) = \hat{P}_n \rho_c, \quad \mathcal{R}(c) = P \rho_c.
$$

定理 14.2.1

如果满足如下条件：

$$
\sup_{c \in \Gamma} \left| \hat{\mathcal{R}}_n(c) - \mathcal{R}(c) \right| \to 0, \quad \mathbb{P}_\theta - \text{a.s.},
$$

则有：

$$\mathcal{R}\left(\hat{\gamma}_n\right) \to \mathcal{R}(\gamma), \quad \mathbb{P}_\theta - \text{a.s.}.$$

证明：

由一致收敛性可以得到：

$$\begin{aligned}
0 &\leq P_\theta\left(\rho_{\hat{\gamma}_n} - \rho_\gamma\right) \\
&= -\left(\hat{P}_n - P_\theta\right)\left(\rho_{\hat{\gamma}_n} - \rho_\gamma\right) + \hat{P}_n\left(\rho_{\hat{\gamma}_n} - \rho_\gamma\right) \\
&\leq -\left(\hat{P}_n - P_\theta\right)\left(\rho_{\hat{\gamma}_n} - \rho_\gamma\right) \\
&\leq \left|\left(\hat{P}_n - P_\theta\right)\rho_{\hat{\gamma}_n}\right| + \left|\left(\hat{P}_n - P_\theta\right)\rho_\gamma\right| \\
&\leq \sup_{c\in\Gamma}\left|\left(\hat{P}_n - P_\theta\right)\rho_c\right| + \left|\left(\hat{P}_n - P_\theta\right)\rho_\gamma\right| \\
&\leq 2\sup_{c\in\Gamma}\left|\left(\hat{P}_n - P_\theta\right)\rho_c\right|.
\end{aligned}$$

需要证明的是，当最小值收敛时，也意味着 $\arg\min$ 的位置收敛.为此，提出以下定义：

定义

如果 $\mathcal{R}(c)$ 的最小化点 γ 满足如下性质，对于所有 $\epsilon > 0$，

$$\inf\{\mathcal{R}(c) : c \in \Gamma, \|c - \gamma\| > \epsilon\} > \mathcal{R}(\gamma)，$$

则称 γ 是良分离的.

若 γ 是良分离的，且 $\mathcal{R}\left(\hat{\gamma}_n\right) \to \mathcal{R}(\gamma), \mathbb{P}_\theta$ -a.s.，则有 $\hat{\gamma}_n \to \gamma, \mathbb{P}_\theta$ -a.s..

引理 14.2.1

若满足以下条件：

1. Γ 是紧致的；
2. $c \mapsto \rho_c(x)$ 对所有 x 连续；
3.

$$P_\theta\left(\sup_{c\in\Gamma}|\rho_c|\right) < \infty.$$

则有一致收敛性：

$$\sup_{c\in\Gamma}\left|\left(\hat{P}_n - P_\theta\right)\rho_c\right| \to 0, \quad \mathbb{P}_\theta - \text{a.s.}.$$

证明：

对每个 $\delta > 0$ 和 $c \in \Gamma$，定义：

$$w(\cdot, \delta, c) := \sup_{\tilde{c}\in\Gamma:\|\tilde{c}-c\|<\delta} |\rho_{\tilde{c}} - \rho_c|.$$

对于所有 x，当 $\delta \downarrow 0$ 时，有：

$$w(x, \delta, c) \to 0.$$

因此，根据支配收敛定理：

$$P_\theta w(\cdot, \delta, c) \to 0.$$

于是，对所有 $\epsilon > 0$，存在 δ_c，使得：

$$P_\theta w(\cdot, \delta_c, c) \leq \epsilon.$$

令：

$$B_c := \{\tilde{c} \in \Gamma : \|\tilde{c} - c\| < \delta_c\}.$$

因为 Γ 是紧致的，$\{B_c : c \in \Gamma\}$ 有有限子覆盖：

$$B_{c_1}, \ldots, B_{c_N}.$$

对 $c \in B_{c_j}$，有：

$$\left|\rho_c - \rho_{c_j}\right| \leq w(\cdot, \delta_{c_j}, c_j).$$

因此：

$$\begin{aligned}\sup_{c\in\Gamma}\left|\left(\hat{P}_n - P_\theta\right)\rho_c\right| &\leq \max_{1\leq j\leq N}\left|\left(\hat{P}_n - P_\theta\right)\rho_{c_j}\right| \\ &\quad + \max_{1\leq j\leq N}\hat{P}_n w(\cdot, \delta_{c_j}, c_j) + \max_{1\leq j\leq N} P_\theta w(\cdot, \delta_{c_j}, c_j) \\ &\to 2\max_{1\leq j\leq N} P_\theta w(\cdot, \delta_{c_j}, c_j) \leq 2\epsilon, \quad \mathbb{P}_\theta - \text{a.s.}.\end{aligned}$$

例子 14.2.1 Logistic 位置家族中 MLE 的一致性

上述定理直接使用了 M 估计量的定义，而不依赖显式表达式.以下是一个无法得到显式表达式的例子.考虑 Logistic 位置家族，其密度函数为：

$$p_\theta(x) = \frac{\mathrm{e}^{x-\theta}}{\left(1 + \mathrm{e}^{x-\theta}\right)^2}, \quad x \in \mathbb{R},$$

其中 $\theta \in \Theta \subset \mathbb{R}$ 是位置参数.定义：

$$\rho_\theta(x) := -\log p_\theta(x) = \theta - x + 2\log\left(1+\mathrm{e}^{x-\theta}\right).$$

因此，$\hat{\theta}_n$ 是最大似然估计量（MLE）.它是以下方程的解：

$$\frac{2}{n}\sum_{i=1}^{n}\frac{\mathrm{e}^{X_i-\hat{\theta}_n}}{1+\mathrm{e}^{X_i-\hat{\theta}_n}} = 1.$$

对于 $\hat{\theta}_n$，我们无法得到其显式表达式.然而，为了应用上述一致性定理，我们需要假设 Θ 是紧致的.这一问题可以通过对 Z 估计量的以下结果加以处理来规避.

定理 14.2.2

假设 $\Gamma \subset \mathbb{R}$，并且对于所有 x，$\psi_c(x)$ 在 c 上连续.此外，假设：

1. $$P_\theta\left|\psi_c\right| < \infty, \quad \forall c;$$
2. 存在 $\delta > 0$，使得：

$$\begin{aligned}\dot{\mathcal{R}}(c) &> 0, \quad \gamma < c < \gamma + \delta,\\ \dot{\mathcal{R}}(c) &< 0, \quad \gamma - \delta < c < \gamma.\end{aligned}$$

那么，对于足够大的 n，$\mathbb{P}_\theta$ 几乎处处存在一个解 $\hat{\gamma}_n$ 使得 $\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0$，且该解 $\hat{\gamma}_n$ 是一致的.

证明

令 $0 < \epsilon < \delta$.根据大数定律，当 n 足够大时，$\mathbb{P}_\theta$ 几乎处处有：

$$\dot{\hat{\mathcal{R}}}_n(\gamma+\epsilon) > 0, \quad \dot{\hat{\mathcal{R}}}_n(\gamma-\epsilon) < 0.$$

由于 $c \mapsto \psi_c$ 的连续性，可得 $\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0$ 对某些满足 $\left|\hat{\gamma}_n - \gamma\right| < \epsilon$ 的 $\hat{\gamma}_n$ 成立.

14.3 M 估计量的渐近正态性

对于一个函数 $f : \mathcal{X} \to \mathbb{R}^p$，定义：

$$P_\theta f := \mathbb{E}_\theta f(X) \in \mathbb{R}^p;$$

若存在，则有：

$$P_\theta f f^T = \mathbb{E}_\theta f(X) f^T(X) \in \mathbb{R}^{p\times p}.$$

向量 $f(X)$ 的协方差矩阵为：

$$\Sigma := P_\theta f f^T - (P_\theta f)(P_\theta f)^T.$$

根据中心极限定理（CLT）：

$$\sqrt{n}\left(\hat{P}_n - P_\theta\right) f \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, \Sigma).$$

按照这种记号，如果 $\{T_n\}$ 是 γ 的渐近线性估计量，则有：

$$T_n - \gamma = \hat{P}_n l_\theta + o_{\mathbb{P}_\theta}(1/\sqrt{n}),$$

其中 $P_\theta l_\theta = 0$，且 $V_\theta := P_\theta l_\theta l_\theta^T < \infty$.

定义：

$$\nu_n(c) := \sqrt{n}\left(\hat{P}_n - P_\theta\right)\psi_c = \sqrt{n}\left(\dot{\hat{\mathcal{R}}}_n(c) - \dot{\mathcal{R}}(c)\right), \quad c \in \Gamma.$$

定理 14.3.1

令 $\hat{\gamma}_n$ 为 γ 的 Z 估计量.假设 $\hat{\gamma}_n$ 是 γ 的一致估计量，并且 ν_n 在 γ 处渐近连续.假设 M_θ^{-1} 存在，且

$$J_\theta := P_\theta \psi_\gamma \psi_\gamma^T,$$

则 $\hat{\gamma}_n$ 是渐近线性的，其影响函数为：

$$l_\theta = -M_\theta^{-1}\psi_\gamma.$$

证明

根据定义：

$$\dot{\hat{\mathcal{R}}}_n\left(\hat{\gamma}_n\right) = 0, \quad \dot{\mathcal{R}}(\gamma) = 0$$

因此，有：

$$\begin{aligned}
0 &= \dot{\hat{\mathcal{R}}}_n\left(\hat{\gamma}_n\right) \\
&= \nu_n\left(\hat{\gamma}_n\right)/\sqrt{n} + \dot{\mathcal{R}}\left(\hat{\gamma}_n\right) \\
&= \nu_n\left(\hat{\gamma}_n\right)/\sqrt{n} + \dot{\mathcal{R}}\left(\hat{\gamma}_n\right) - \dot{\mathcal{R}}(\gamma) \\
&= (i) + (ii)
\end{aligned}$$

对于第一个项，利用 ν_n 在 γ 处的渐近连续性：

$$\begin{aligned}
(i) &= \nu_n\left(\hat{\gamma}_n\right)/\sqrt{n} \\
&= \nu_n(\gamma)/\sqrt{n} + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \\
&= \dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n})
\end{aligned}$$

对于第二个项，利用 $\dot{\mathcal{R}}(c) = P_\theta \psi_c$ 在 $c = \gamma$ 的可微性：

$$\begin{aligned}
(ii) &= \dot{\mathcal{R}}\left(\hat{\gamma}_n\right) - \dot{\mathcal{R}}(\gamma) \\
&= M_\theta\left(\hat{\gamma}_n - \gamma\right) + o\left(\left\|\hat{\gamma}_n - \gamma\right\|\right)
\end{aligned}$$

于是有：

$$0 = \dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) + M_\theta\left(\hat{\gamma}_n - \gamma\right) + o\left(\left\|\hat{\gamma}_n - \gamma\right\|\right)$$

由于根据中心极限定理（CLT），$\dot{\hat{\mathcal{R}}}_n(\gamma) = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$，这表明 $\left\|\hat{\gamma}_n - \gamma\right\| = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$.因此：

$$0 = \dot{\hat{\mathcal{R}}}_n(\gamma) + M_\theta\left(\hat{\gamma}_n - \gamma\right) + o_{\mathbf{P}_\theta}(1/\sqrt{n})$$

或者：

$$\begin{aligned} M_\theta\left(\hat{\gamma}_n - \gamma\right) &= -\dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \\ &= -\hat{P}_n\psi_\gamma + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \end{aligned}$$

因此：

$$\left(\hat{\gamma}_n - \gamma\right) = -M_\theta^{-1}\hat{P}_n\psi_\gamma + o_{\mathbf{P}_\theta}(1/\sqrt{n})$$

推论 14.3.1

在定理 14.3.1 的条件下，有：

$$\sqrt{n}\left(\hat{\gamma}_n - \gamma\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, V_\theta\right)$$

其中：

$$V_\theta = M_\theta^{-1} J_\theta M_\theta^{-1}$$

定理 14.3.2

在一些更严格的条件下，可以直接证明渐近线性.假设 $\hat{\gamma}_n$ 是 γ 的一致估计量.假设对于 γ 的某个邻域 $\{c \in \Gamma : \|c - \gamma\| < \epsilon\}$ 内的所有 c，映射 $c \mapsto \psi_c(x)$ 对所有 x 可微，其导数为：

$$\dot{\psi}_c(x) = \frac{\partial}{\partial c^T}\psi_c(x)$$

这是一个 $p \times p$ 的矩阵.此外，假设对于 γ 邻域内的所有 c 和 $\tilde{c}$ 以及所有 x，在矩阵范数意义下满足：

$$\left\|\dot{\psi}_c(x) - \dot{\psi}_{\tilde{c}}(x)\right\| \leq H(x)\|c - \tilde{c}\|$$

其中 $H : \mathcal{X} \to \mathbb{R}$ 满足：

$$P_\theta H < \infty$$

则有：

$$M_\theta := \frac{\partial}{\partial c^T}\dot{\mathcal{R}}(c)\Big|_{c=\gamma} = P_\theta\dot{\psi}_\gamma$$

若 M_θ^{-1} 和 $J_\theta := \mathbb{E}_\theta\psi_\gamma\psi_\gamma^T$ 存在，则 $\hat{\gamma}_n$ 的影响函数为：

$$l_\theta = -M_\theta^{-1}\psi_\gamma$$

证明

(14.3) 的推导

由支配收敛定理，(14.3) 成立.根据均值定理：

$$\begin{aligned}0 &= \dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) \\ &= \hat{P}_n \psi_{\hat{\gamma}_n} \\ &= \hat{P}_n \psi_\gamma + \hat{P}_n \dot{\psi}_{\tilde{\gamma}_n(\cdot)}(\hat{\gamma}_n - \gamma) \\ &= \dot{\hat{\mathcal{R}}}_n(\gamma) + \hat{P}_n \dot{\psi}_{\tilde{\gamma}_n(\cdot)}(\hat{\gamma}_n - \gamma)\end{aligned}$$

其中 $\tilde{\gamma}_n(x)$ 满足 $\|\tilde{\gamma}_n(x) - \gamma\| \le \|\hat{\gamma}_n - \gamma\|$.因此：

$$0 = \dot{\hat{\mathcal{R}}}_n(\gamma) + \hat{P}_n \dot{\psi}_\gamma(\hat{\gamma}_n - \gamma) + \hat{P}_n(\dot{\psi}_{\tilde{\gamma}_n(\cdot)} - \dot{\psi}_\gamma)(\hat{\gamma}_n - \gamma)$$

于是有：

$$\left\| \dot{\hat{\mathcal{R}}}_n(\gamma) + \hat{P}_n \dot{\psi}_\gamma(\hat{\gamma}_n - \gamma) \right\| \le (\hat{P}_n H)\|\hat{\gamma}_n - \gamma\|^2 = \mathcal{O}_{\mathbf{P}_\theta}(1)\|\hat{\gamma}_n - \gamma\|^2$$

其中使用了 $P_\theta H < \infty$.根据大数定律，有：

$$\hat{P}_n \dot{\psi}_\gamma = P_\theta \dot{\psi}_\gamma + o_{\mathbf{P}_\theta}(1) = M_\theta + o_{\mathbf{P}_\theta}(1)$$

因此：

$$\left| \dot{\hat{\mathcal{R}}}_n(\gamma) + M_\theta(\hat{\gamma}_n - \gamma) + o_{\mathbf{P}_\theta}(\|\hat{\gamma}_n - \gamma\|) \right| = \mathcal{O}_{\mathbf{P}_\theta}(\|\hat{\gamma}_n - \gamma\|^2)$$

由中心极限定理（CLT），$\dot{\hat{\mathcal{R}}}_n(\gamma) = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$，这确保 $\|\hat{\gamma}_n - \gamma\| = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$.因此：

$$\left| \dot{\hat{\mathcal{R}}}_n(\gamma) + M_\theta(\hat{\gamma}_n - \gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \right| = \mathcal{O}_{\mathbf{P}_\theta}(1/n)$$

于是有：

$$M_\theta(\hat{\gamma}_n - \gamma) = -\dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n})$$

因此：

$$\begin{aligned}\hat{\gamma}_n - \gamma &= -M_\theta^{-1}\dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \\ &= -\hat{P}_n M_\theta^{-1}\psi_\gamma + o_{\mathbf{P}_\theta}(1/\sqrt{n})\end{aligned}$$

最大似然估计的渐近正态性

在满足正则性条件下，最大似然估计（MLE）是渐近正态的，其渐近协方差矩阵是费舍尔信息矩阵的逆.设 $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$，对支配测度 ν 的密度为 $p_\theta = dP_\theta/d\nu$，$\Theta \subset \mathbb{R}^p$.假设 p_θ 的支持集不依赖于 θ，并以负对数密度为损失函数：

$$\rho_\theta := -\log p_\theta$$

MLE 的定义为：

$$\hat{\theta}_n := \arg\max_{\tilde{\theta}\in\Theta} \hat{P}_n \log p_{\tilde{\theta}}$$

假设存在得分函数：

$$s_\theta = \frac{\partial}{\partial\theta}\log p_\theta = \frac{\dot{p}_\theta}{p_\theta}$$

并且得分函数的期望为零：

$$P_\theta s_\theta = \int \dot{p}_\theta d\nu = 0$$

费舍尔信息矩阵定义为：

$$I(\theta) := P_\theta s_\theta s_\theta^T$$

进一步可得：

$$M_\theta = P_\theta \dot{\psi}_\theta = -P_\theta \dot{s}_\theta$$

且：

$$P_\theta \dot{s}_\theta = -I(\theta)$$

因此 $M_\theta = I(\theta)$，MLE 的影响函数为：

$$l_\theta = I(\theta)^{-1} s_\theta$$

MLE 的渐近协方差矩阵为：

$$I(\theta)^{-1}\left(P_\theta s_\theta s_\theta^T\right) I(\theta)^{-1} = I(\theta)^{-1}$$

最终得到：

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, I^{-1}(\theta))$$

14.5 两个 M-估计的进一步例子

在本节中，我们考察 α 分位数和 Huber 估计量的渐近性质.

例子 14.5.1 α 分位数的渐近正态性

在这个例子中，感兴趣的参数是 α 分位数.我们采用一个不满足正则性条件的损失函数，但它仍然导出一个渐近线性（因此渐近正态）的估计量.

设 $\mathcal{X} := \mathbb{R}$，$X$ 的分布函数记为 F.给定 $0 < \alpha < 1$，定义 α 分位数为 $\gamma = F^{-1}(\alpha)$（假设存在）. 此外，假设 F 对勒贝格测度的密度为 f，且 $f(x) > 0$ 在 γ 的某个邻域内成立.我们选取以下损

失函数：

$$\rho_c(x) := \rho(x-c),$$

其中

$$\rho(x) := (1-\alpha)|x| \cdot 1\{x<0\} + \alpha|x| \cdot 1\{x>0\}.$$

验证 $\mathcal{R}(c) := P_\theta \rho_c$ 的最小值位置为 γ：

$$\arg\min_c \mathcal{R}(c) = F^{-1}(\alpha) := \gamma.$$

由 $\dot{\rho}(x)$ 的定义可得：

$$\dot{\rho}(x) = \alpha \cdot 1\{x>0\} - (1-\alpha)\cdot 1\{x<0\}.$$

$\dot{\rho}(x)$ 在 $x=0$ 不连续，这是此例中的一个不规则性.对 $\psi_c(x)$ 有：

$$\psi_c(x) = -\alpha \cdot 1\{x>c\} + (1-\alpha)\cdot 1\{x<c\}.$$

计算 $\dot{\mathcal{R}}(c)$：

$$\dot{\mathcal{R}}(c) = P_\theta \psi_c = -\alpha + F(c).$$

从而 $\dot{\mathcal{R}}(\gamma)=0$，且 $\gamma = F^{-1}(\alpha)$.

渐近影响函数：

计算 M_θ：

$$M_\theta = \frac{d}{dc}\dot{\mathcal{R}}(c)\Big|_{c=\gamma} = f(\gamma) = f(F^{-1}(\alpha)).$$

影响函数为：

$$l_\theta(x) = -M_\theta^{-1}\psi_\gamma(x) = \frac{1}{f(\gamma)}\{-1\{x<\gamma\} + \alpha\}.$$

最终结论：

$$\sqrt{n}\left(\hat{F}_n^{-1}(\alpha) - F^{-1}(\alpha)\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}\right).$$

例子 14.5.2 Huber 估计量的渐近正态性

我们证明 Huber 估计量是渐近线性的，因此是渐近正态的.令 $\mathcal{X}=\mathbb{R}$，F 为 X 的分布函数.目标参数为位置参数，Huber 损失函数为：

$$\rho_c(x) = \rho(x-c),$$

其中

$$\rho(x)=\begin{cases}x^2 & |x|\le k\\ k(2|x|-k) & |x|>k\end{cases}.$$

定义：

$$\gamma:=\arg\min_{c} P_\theta\rho_c.$$

一阶导数 $\psi_c(x)$：

$$\dot\rho(x)=\begin{cases}2x & |x|\le k\\ +2k & x>k\\ -2k & x<-k\end{cases},$$

$$\psi_c(x)=\begin{cases}-2(x-c) & |x-c|\le k\\ -2k & x-c>k\\ +2k & x-c<-k\end{cases}.$$

对 $\dot{\mathcal{R}}(c)$ 有：

$$\begin{aligned}\dot{\mathcal{R}}(c)=&-2\int_{-k+c}^{k+c}xdF(x)+2c[F(k+c)-F(-k+c)]\\&-2k[1-F(k+c)]+2kF(-k+c).\end{aligned}$$

计算 M_θ：

$$M_\theta=\frac{d}{dc}\dot{\mathcal{R}}(c)\Big|_{c=\gamma}=2[F(k+\gamma)-F(-k+\gamma)].$$

渐近影响函数：

$$l_\theta(x)=\frac{1}{[F(k+\gamma)-F(-k+\gamma)]}\begin{cases}x-\gamma & |x-\gamma|\le k\\ +k & x-\gamma>k\\ -k & x-\gamma<-k\end{cases}.$$

当 $k\to 0$ 时，这与中位数的影响函数一致.

14.6 渐近相对效率

在本节中，假设感兴趣的参数是实值的：

$$\gamma\in\Gamma\subset\mathbb{R}$$

定义 14.6.1

设 $T_{n,1}$ 和 $T_{n,2}$ 是 γ 的两个估计量，满足

$$\sqrt{n}\left(T_{n,j}-\gamma\right)\xrightarrow{\mathcal{D}_\theta}\mathcal{N}\left(0,V_{\theta,j}\right),\quad j=1,2.$$

则

$$\mathrm{e}_{2:1}:=\frac{V_{\theta,1}}{V_{\theta,2}}$$

称为 $T_{n,2}$ 相对于 $T_{n,1}$ 的渐近相对效率.

- 若 $\mathrm{e}_{2:1}>1$，则 $T_{n,2}$ 渐近上比 $T_{n,1}$ 更高效.
- 基于 $T_{n,2}$ 的渐近 $(1-\alpha)$ 置信区间比基于 $T_{n,1}$ 的更窄.

例子 14.6.1 样本均值与样本中位数的渐近相对效率

设 $\mathcal{X}=\mathbb{R}$，F 是 X 的分布函数.假设 F 关于 μ 对称，即

$$F(\cdot)=F_0(\cdot-\mu),$$

其中 F_0 关于零对称.假设 F_0 具有有限方差 σ^2，且对勒贝格测度的密度为 f_0，满足 $f_0(0)>0$.取 $T_{n,1}:=\bar{X}_n$（样本均值）和 $T_{n,2}:=\hat{F}_n^{-1}(1/2)$（样本中位数）.

- $T_{n,1}$ 的渐近方差为 $V_{\theta,1}=\sigma^2$.
- $T_{n,2}$ 的渐近方差为 $V_{\theta,2}=1/(4f_0^2(0))$（见例子 14.5.1）.

因此，渐近相对效率为：

$$\mathrm{e}_{2:1}=4\sigma^2 f_0^2(0).$$

效率的高低取决于 F_0 的具体分布：

1. **情况 i**：若 F_0 为标准正态分布，即 $F_0=\Phi$.
   - $\sigma^2=1$，$f_0(0)=1/\sqrt{2\pi}$.
   - 则
   $$\mathrm{e}_{2:1}=\frac{2}{\pi}\approx 0.64$$
   样本均值 $\bar{X}_n$ 更高效.
2. **情况 ii**：若 F_0 为 Laplace 分布，且方差 $\sigma^2=1$.
   - 密度函数为
   $$f_0(x)=\frac{1}{\sqrt{2}}\exp[-\sqrt{2}|x|],\quad x\in\mathbb{R}.$$
   - 则 $f_0(0)=1/\sqrt{2}$，因此
   $$\mathrm{e}_{2:1}=2$$
   样本中位数 $\hat{F}_n^{-1}(1/2)$ 更高效.
3. **情况 iii**：若 F_0 是如下混合分布：
   $$F_0=(1-\eta)\Phi+\eta\Phi(\cdot/3),$$

表示 X 的分布是单位方差的正态分布 $\mathcal{N}(\mu,1)$ 和方差为 9 的正态分布 $\mathcal{N}(\mu,9)$ 的混合分布，其中混合比例为 $1-\eta$ 和 η.

- X 的方差为：

$$\sigma^2=(1-\eta)\cdot 1+\eta\cdot 9=1+8\eta.$$

- $f_0(0)$ 为：

$$f_0(0)=\frac{1}{\sqrt{2\pi}}\left(1-\frac{2\eta}{3}\right).$$

- 渐近相对效率为：

$$\mathrm{e}_{2:1}=\frac{2}{\pi}\left(1-\frac{2\eta}{3}\right)^2(1+8\eta).$$

α -截尾均值的渐近相对效率

对于对称分布 F，α -截尾均值与 Huber 估计量有相同的影响函数，截尾边界为 $k=F^{-1}(1-\alpha)$：

$$l_\theta(x)=\frac{1}{F_0(k)-F_0(-k)}\begin{cases}x-\mu, & |x-\mu|\le k\\ +k, & x-\mu>k\\ -k, & x-\mu<-k\end{cases}.$$

渐近方差：

$$V_{\theta,\alpha}=\frac{\int_{F_0^{-1}(\alpha)}^{F_0^{-1}(1-\alpha)}x^2dF_0(x)+2\alpha\left(F_0^{-1}(1-\alpha)\right)^2}{(1-2\alpha)^2}.$$

从而计算 α -截尾均值相对于均值的渐近相对效率：

表：α -截尾均值相对效率

α 截尾比例	$\alpha=0.05$	0.125	0.5（中位数）
$\eta=0.00$	0.99	0.94	0.64
$\eta=0.05$	1.20	1.19	0.83
$\eta=0.25$	1.40	1.66	1.33

14.7 渐近枢轴量

在本节中，假设具有足够的正则性条件，如导数的存在及积分和微分的交换.

渐近枢轴量的定义

回忆渐近枢轴量的定义（见第 6.2 节）.它是关于数据 $X_1, \ldots, X_n$ 和感兴趣参数 $\gamma = g(\theta) \in \mathbb{R}^p$ 的函数 $Z_n(\gamma) := Z_n\left(X_1, \ldots, X_n, \gamma\right)$，满足其渐近分布不依赖于未知参数 θ，即：

$$Z_n(\gamma) \xrightarrow{\mathcal{D}_\theta} Z, \forall \theta,$$

其中，Z 是分布为 Q 的随机变量，且 Q 不依赖于 θ.

渐近枢轴量可用于构造参数 γ 的近似 $(1-\alpha)$ 置信区间以及近似显著性水平为 α 的假设检验 $H_0 : \gamma = \gamma_0$.

渐近正态估计量的枢轴量构造

考虑一个渐近正态估计量 T_n，其渐近无偏，且渐近协方差矩阵为 V_θ，即

$$\sqrt{n}\left(T_n - \gamma\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, V_\theta\right), \forall \theta,$$

假设这样的估计量存在.

第一类渐近枢轴量

若渐近协方差矩阵 V_θ 可逆，且仅依赖于感兴趣的参数 γ，例如 $V_\theta = V(\gamma)$，则渐近枢轴量为：

$$Z_{n,1}(\gamma) := n(T_n - \gamma)^T V(\gamma)^{-1}\left(T_n - \gamma\right)$$

其渐近分布为自由度为 p 的 χ^2 分布.

第二类渐近枢轴量

若存在一致估计量 $\hat{V}_n$ 估计 V_θ，则渐近枢轴量为：

$$Z_{n,2}(\gamma) := n(T_n - \gamma)^T \hat{V}_n^{-1}\left(T_n - \gamma\right).$$

其渐近分布同样为自由度为 p 的 χ^2 分布.这一结果来自于 Slutsky 引理.

渐近方差的估计

1. 若 $\hat{\theta}_n$ 是 θ 的一致估计量，且 $\theta \mapsto V_\theta$ 连续，可令 $\hat{V}_n := V_{\hat{\theta}_n}$.

2. 若 $T_n = \hat{\gamma}_n$ 是 γ 的 M 估计量，且 γ 是方程 $P_\theta \psi_\gamma = 0$ 的解，则（在正则性条件下）渐近协方差矩阵为：

$$V_\theta = M_\theta^{-1} J_\theta M_\theta^{-1},$$

其中：

$$J_\theta = P_\theta \psi_\gamma \psi_\gamma^T,$$

$$M_\theta = \ddot{\mathcal{R}}(\gamma) = P_\theta \dot{\psi}_\gamma.$$

对 J_θ 和 M_θ 的估计分别为：

$$\hat{J}_n := \hat{P}_n \psi_{\hat{\gamma}_n} \psi_{\hat{\gamma}_n}^T = \frac{1}{n} \sum_{i=1}^n \psi_{\hat{\gamma}_n}(X_i) \psi_{\hat{\gamma}_n}^T(X_i),$$

$$\hat{M}_n := \ddot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = \hat{P}_n \dot{\psi}_{\hat{\gamma}_n} = \frac{1}{n} \sum_{i=1}^n \dot{\psi}_{\hat{\gamma}_n}(X_i).$$

在某些正则性条件下，

$$\hat{V}_n := \hat{M}_n^{-1} \hat{J}_n \hat{M}_n^{-1}$$

是一致估计量.

14.8 基于 MLE 的渐近枢轴量

假设 $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$，其中 $\Theta \subset \mathbb{R}^p$，且 $\mathcal{P}$ 被某 σ -有限测度 ν 控制.令 $p_\theta := dP_\theta / d\nu$ 表示密度函数.MLE 定义为：

$$\hat{\theta}_n := \arg\max_{\vartheta \in \Theta} \sum_{i=1}^n \log p_\vartheta(X_i).$$

在正则性条件下，$\hat{\theta}_n$ 是带有损失函数 $\rho_\vartheta = -\log p_\vartheta$ 的 M 估计量.因此，$\psi_\vartheta = \dot{\rho}_\vartheta$ 是负的得分函数：

$$s_\vartheta := \frac{\dot{p}_\vartheta}{p_\vartheta}.$$

MLE 的渐近协方差矩阵为 Fisher 信息矩阵的逆：

$$I(\theta) := P_\theta s_\theta s_\theta^T,$$

其渐近正态性为：

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, I^{-1}(\theta)\right), \forall \theta.$$

因此，第一类渐近枢轴量为：

$$Z_{n,1}(\theta)=n\left(\hat{\theta}_n-\theta\right)^T I(\theta)\left(\hat{\theta}_n-\theta\right),$$

若 $\hat{I}_n$ 是 $I(\theta)$ 的一致估计量，则第二类渐近枢轴量为：

$$Z_{n,2}(\theta)=n\left(\hat{\theta}_n-\theta\right)^T \hat{I}_n\left(\hat{\theta}_n-\theta\right).$$

其中 Fisher 信息的估计量为：

$$\hat{I}_n:=-\frac{1}{n}\sum_{i=1}^n \dot{s}_{\hat{\theta}_n}(X_i)=-\frac{\partial^2}{\partial\vartheta\partial\vartheta^T}\frac{1}{n}\sum_{i=1}^n \log p_\vartheta(X_i)\Big|_{\vartheta=\hat{\theta}_n}.$$

14.7 第三类渐近枢轴量

定义对数似然比

定义二倍对数似然比为：

$$2\mathcal{L}_n\left(\hat{\theta}_n\right)-2\mathcal{L}_n(\theta):=2\sum_{i=1}^n\left[\log p_{\hat{\theta}_n}(X_i)-\log p_\theta(X_i)\right]$$

对数似然比实际上是一个渐近枢轴量，其实用优势在于它是**自归一化**的：不需要显式地估计渐近协方差.

引理 14.8.1

在正则性条件下，$2\mathcal{L}_n\left(\hat{\theta}_n\right)-2\mathcal{L}_n(\theta)$ 是关于 θ 的渐近枢轴量，其渐近分布为自由度为 p 的 χ^2 分布：

$$2\mathcal{L}_n\left(\hat{\theta}_n\right)-2\mathcal{L}_n(\theta)\xrightarrow{\mathcal{D}_\theta}\chi_p^2\quad\forall\theta.$$

证明思路

通过两项泰勒展开有：

$$\begin{aligned}2\mathcal{L}_n\left(\hat{\theta}_n\right)-2\mathcal{L}_n(\theta)&=2n\hat{P}_n\left[\log p_{\hat{\theta}_n}-\log p_\theta\right]\\&\approx 2n\left(\hat{\theta}_n-\theta\right)^T\hat{P}_n s_\theta+n\left(\hat{\theta}_n-\theta\right)^T\hat{P}_n\dot{s}_\theta\left(\hat{\theta}_n-\theta\right)\\&\approx 2n\left(\hat{\theta}_n-\theta\right)^T\hat{P}_n s_\theta-n\left(\hat{\theta}_n-\theta\right)^T I(\theta)\left(\hat{\theta}_n-\theta\right),\end{aligned}$$

其中第二步使用了 $\hat{P}_n\dot{s}_\theta\approx P_\theta\dot{s}_\theta=-I(\theta)$.

注意，MLE $\hat{\theta}_n$ 是渐近线性的，其影响函数为 $l_\theta=I(\theta)^{-1}s_\theta$，即：

$$\hat{\theta}_n - \theta = I(\theta)^{-1}\hat{P}_n s_\theta + o_{\mathbf{P}_\theta}\left(n^{-1/2}\right).$$

因此有：

$$2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n(\theta) \approx n\left(\hat{P}_n s_\theta\right)^T I(\theta)^{-1}\left(\hat{P}_n s_\theta\right).$$

由渐近正态性：

$$\sqrt{n}\hat{P}_n s_\theta \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, I(\theta)),$$

可得结论.

14.9 多项分布的 MLE

假设 $X_1, \ldots, X_n$ 是 i.i.d. 样本，其中 $X \in \{1, \ldots, k\}$ 是一个标签，且

$$P_\theta(X = j) := \pi_j, \quad j = 1, \ldots, k,$$

其中概率 π_j 满足 $\pi_j > 0$ 且 $\sum_{j=1}^k \pi_j = 1$.定义未知参数为 $\theta = (\pi_1, \ldots, \pi_{k-1})$，即有 $p := k - 1$ 个未知参数.

引理 14.9.1

对于每个 $j = 1, \ldots, k$，π_j 的 MLE 为：

$$\hat{\pi}_j = \frac{N_j}{n},$$

其中 $N_j := \#\{i : X_i = j\}$ 表示观测值等于 j 的样本个数.

证明

对数密度函数为：

$$\log p_\theta(x) = \sum_{j=1}^k 1\{x = j\} \log \pi_j,$$

因此对数似然函数为：

$$\sum_{i=1}^n \log p_\theta(X_i) = \sum_{j=1}^k N_j \log \pi_j.$$

对参数 $\theta = (\pi_1, \ldots, \pi_{k-1})$ 求导（注意 $\pi_k = 1 - \sum_{j=1}^{k-1} \pi_j$），令导数为零：

$$\frac{N_j}{\hat{\pi}_j} - \frac{N_k}{\hat{\pi}_k} = 0.$$

因此有：

$$\hat{\pi}_j = N_j \frac{\hat{\pi}_k}{N_k}, \quad j = 1, \ldots, k.$$

由于 $\sum_{j=1}^{k} \hat{\pi}_j = 1$，可得：

$$\hat{\pi}_k = \frac{N_k}{n},$$

进而：

$$\hat{\pi}_j = \frac{N_j}{n}, \quad j = 1, \ldots, k.$$

Fisher 信息

定义 Fisher 信息矩阵 $I(\theta)$ 为：

$$I(\theta) = \begin{pmatrix} \frac{1}{\pi_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\pi_{k-1}} \end{pmatrix} + \frac{1}{\pi_k} \iota \iota^T,$$

其中 ι 是 $(k-1)$ 维向量，$\iota := (1, \ldots, 1)^T$.

证明

根据定义，得：

$$s_{\theta,j}(x) = \frac{1}{\pi_j} \mathbf{1}\{x = j\} - \frac{1}{\pi_k} \mathbf{1}\{x = k\}$$

因此 Fisher 信息矩阵的 (j_1, j_2) 元素为：

$$\begin{aligned} (I(\theta))_{j_1, j_2} &= \mathbb{E}_\theta \left(\frac{1}{\pi_{j_1}} \mathbf{1}\{X = j_1\} - \frac{1}{\pi_k} \mathbf{1}\{X = k\} \right) \left(\frac{1}{\pi_{j_2}} \mathbf{1}\{X = j_2\} - \frac{1}{\pi_k} \mathbf{1}\{X = k\} \right) \\ &= \begin{cases} \frac{1}{\pi_k}, & j_1 \neq j_2, \\ \frac{1}{\pi_j} + \frac{1}{\pi_k}, & j_1 = j_2 = j. \end{cases} \end{aligned}$$

于是 Fisher 信息矩阵为：

$$I(\theta) = \text{对角线元素为 } \frac{1}{\pi_j}, \text{ 其余元素为 } \frac{1}{\pi_k}.$$

计算渐近枢轴量 $Z_{n,1}(\theta)$

$$
\begin{aligned}
Z_{n,1}(\theta) &= n\left(\hat{\theta}_n - \theta\right)^T I(\theta) \left(\hat{\theta}_n - \theta\right) \\
&= n \begin{pmatrix} \hat{\pi}_1 - \pi_1 \\ \vdots \\ \hat{\pi}_{k-1} - \pi_{k-1} \end{pmatrix}^T \left[\begin{pmatrix} \frac{1}{\pi_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\pi_{k-1}} \end{pmatrix} + \frac{1}{\pi_k} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \right] \\
&\quad \times \begin{pmatrix} \hat{\pi}_1 - \pi_1 \\ \vdots \\ \hat{\pi}_{k-1} - \pi_{k-1} \end{pmatrix} \\
&= n \sum_{j=1}^{k-1} \frac{\left(\hat{\pi}_j - \pi_j\right)^2}{\pi_j} + n \frac{1}{\pi_k} \left(\sum_{j=1}^{k-1} \left(\hat{\pi}_j - \pi_j\right) \right)^2 .
\end{aligned}
$$

由 $\hat{\pi}_k = 1 - \sum_{j=1}^{k-1} \hat{\pi}_j$，可以简化为：

$$
Z_{n,1}(\theta) = n \sum_{j=1}^{k} \frac{\left(\hat{\pi}_j - \pi_j\right)^2}{\pi_j} = \sum_{j=1}^{k} \frac{\left(N_j - n\pi_j\right)^2}{n\pi_j}.
$$

这即为皮尔逊卡方统计量的形式：

$$
\sum \frac{(\text{观察值} - \text{期望值})^2}{\text{期望值}}
$$

渐近枢轴量 $Z_{n,2}(\theta)$

若将 Fisher 信息矩阵中的 π_j 替换为其估计值 $\hat{\pi}_j$，则得到渐近枢轴量：

$$
Z_{n,2}(\theta) = \sum_{j=1}^{k} \frac{\left(N_j - n\pi_j\right)^2}{N_j}.
$$

这即为皮尔逊卡方统计量的另一种形式：

$$
\sum \frac{(\text{观察值} - \text{期望值})^2}{\text{观察值}}
$$

对数似然比渐近枢轴量

对数似然比渐近枢轴量为：

$$
2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n(\theta) = 2 \sum_{j=1}^{k} N_j \log\left(\frac{\hat{\pi}_j}{\pi_j}\right).
$$

利用近似 $\log(1+x) \approx x - x^2/2$，可以得到：

$$2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n(\theta) \approx Z_{n,2}(\theta).$$

三种渐近枢轴量的分布

三种渐近枢轴量 $Z_{n,1}(\theta)$、$Z_{n,2}(\theta)$ 和 $2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n(\theta)$ 在 H_0 下均满足渐近分布为 χ^2_{k-1}.

14.10 似然比检验

对于简单假设 $H_0 : \theta = \theta_0$，可以使用对数似然比统计量：

$$2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n\left(\theta_0\right)$$

作为检验统计量.拒绝 H_0 的条件为：

$$2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n\left(\theta_0\right) > G_p^{-1}(1-\alpha),$$

其中 G_p 是自由度为 p 的 χ^2 分布的分布函数.

对于约束假设 $H_0 : R(\theta) = 0$，对数似然比为：

$$2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n\left(\hat{\theta}_n^0\right),$$

其中 $\hat{\theta}_n$ 是无约束 MLE，$\hat{\theta}_n^0$ 是约束 MLE.

引理 14.10.1：在正则性条件下，若 $H_0 : R(\theta) = 0$ 成立，则有：

$$2\mathcal{L}_n\left(\hat{\theta}_n\right) - 2\mathcal{L}_n\left(\hat{\theta}_n^0\right) \xrightarrow{\mathcal{D}} \chi^2_q.$$

证明概要

定义随机变量：

$$\mathbf{Z}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_\theta(X_i)$$

与引理 14.8.1 的证明类似，我们可以使用二阶泰勒展开，对任意满足 $\vartheta_n = \theta + \mathcal{O}_{\mathbf{P}_\theta}(n^{-1/2})$ 的序列 ϑ_n，有：

$$2\sum_{i=1}^{n} \left[\log p_{\vartheta_n}(X_i) - \log p_\theta(X_i)\right]$$

$$= 2\sqrt{n}(\vartheta_n - \theta)^T \mathbf{Z}_n - n(\vartheta_n - \theta)^T I(\theta)(\vartheta_n - \theta) + o_{\mathbf{P}_\theta}(1)$$

这里我们用到了 $\sum_{i=1}^{n} \dot{s}_{\vartheta_n}(X_i)/n = -I(\theta) + o_{\mathbf{P}_\theta}(1)$.

此外，通过一阶泰勒展开，并利用 $R(\theta) = 0$，得：

$$R(\vartheta_n) = \dot{R}(\theta)(\vartheta_n - \theta) + o_{\mathbf{P}_\theta}(n^{-1/2})$$

结合引理 12.4.1 的结果，令 $z := \mathbf{Z}_n, B := \dot{R}(\theta), V = I(\theta)$，得到：

$$\begin{aligned} 2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}_n^0) = \mathbf{Z}_n^T I(\theta)^{-1}\dot{R}(\theta)^T \Big(\dot{R}(\theta)I(\theta)^{-1}\dot{R}(\theta)^T\Big)^{-1} \dot{R}(\theta)I(\theta)^{-1}\mathbf{Z}_n \\ + o_{\mathbf{P}_\theta}(1) \\ := \mathbf{Y}_n^T W^{-1}\mathbf{Y}_n + o_{\mathbf{P}_\theta}(1) \end{aligned}$$

其中，$\mathbf{Y}_n$ 是 q 维向量：

$$\mathbf{Y}_n := \dot{R}(\theta)I(\theta)^{-1}\mathbf{Z}_n$$

W 是 $(q \times q)$ 矩阵：

$$W := \dot{R}(\theta)I(\theta)^{-1}\dot{R}(\theta)^T$$

由 $\mathbf{Z}_n \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, I(\theta))$，可得：

$$\mathbf{Y}_n \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, W)$$

因此，

$$\mathbf{Y}_n^T W^{-1}\mathbf{Y}_n \xrightarrow{\mathcal{D}_\theta} \chi_q^2$$

推论 14.10.1

由引理 14.10.1 的证明概要，可以进一步推得：

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}_n^0) \approx n(\hat{\theta}_n - \hat{\theta}_n^0)^T I(\theta)(\hat{\theta}_n - \hat{\theta}_n^0)$$

以及：

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}_n^0) \approx n(\hat{\theta}_n - \hat{\theta}_n^0)^T I(\hat{\theta}_n^0)(\hat{\theta}_n - \hat{\theta}_n^0)$$

14.11 列联表

假设 X 是一个双变量标签，取值范围为 $\{(j,k) : j = 1, \ldots, r, k = 1, \ldots, s\}$.例如，第一个索引可以对应性别 $(r = 2)$，第二个索引可以对应眼睛颜色 $(s = 3)$.令组合 (j,k) 的概率为：

$$\pi_{j,k} := P_\theta(X = (j,k))$$

设 $X_1, \ldots, X_n$ 为 X 的独立同分布样本，定义：

$$N_{j,k} := \#\{X_i = (j,k)\}$$

根据 14.9 节的结果，无约束情况下 $\pi_{j,k}$ 的 MLE 为：

$$\hat{\pi}_{j,k} := \frac{N_{j,k}}{n}$$

我们希望检验两个标签是否独立.零假设为：

$$H_0 : \pi_{j,k} = (\pi_{j,+}) \times (\pi_{+,k}), \quad \forall (j,k)$$

其中：

$$\pi_{j,+} := \sum_{k=1}^{s} \pi_{j,k}, \quad \pi_{+,k} := \sum_{j=1}^{r} \pi_{j,k}$$

可以验证，受约束的 MLE 为：

$$\hat{\pi}^0_{j,k} = (\hat{\pi}_{j,+}) \times (\hat{\pi}_{+,k}),$$

其中：

$$\hat{\pi}_{j,+} := \sum_{k=1}^{s} \hat{\pi}_{j,k}, \quad \hat{\pi}_{+,k} := \sum_{j=1}^{r} \hat{\pi}_{j,k}$$

对数似然比检验统计量

对数似然比检验统计量为：

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}^0_n) = 2 \sum_{j=1}^{r} \sum_{k=1}^{s} N_{j,k} \log \left(\frac{n N_{j,k}}{N_{j,+} N_{+,k}} \right)$$

其近似表达式为：

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}^0_n) \approx n \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(N_{j,k} - N_{j,+} N_{+,k}/n)^2}{N_{j,+} N_{+,k}}$$

这正是检验独立性的皮尔逊卡方检验统计量.

在该例中，自由度 q 为：

$$q = (rs - 1) - ((r-1) + (s-1)) = (r-1)(s-1)$$

因此，在 H_0 下，有：

$$n \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(N_{j,k} - N_{j,+} N_{+,k}/n)^2}{N_{j,+} N_{+,k}} \xrightarrow{\mathcal{D}_\theta} \chi^2_{(r-1)(s-1)}$$

第十五章

抽象渐近分析 ⋆

在 2.4.1 小节中，我们讨论了所谓的**替代估计量**（plug-in estimators）.这一思想是将估计量 $\hat{\gamma}_n$ 通常表示为经验分布 $\hat{P}_n$ 的某个泛函 Q.当 P 为"真实"分布且 $\gamma = Q(P)$ 时，研究的核心是当 $\hat{P}_n$ 接近 P 时，$\hat{\gamma}_n = Q(\hat{P}_n)$ 与 $\gamma = Q(P)$ 的接近程度.本章第一部分探讨这一主题.

在 5.5 节中，我们得到了Cramér-Rao **下界**.然而，一个略显令人失望的结果是，仅在指数族中才能达到此下界（见引理 5.6.1）.我们还在 14.4 节中看到，MLE 渐近无偏并渐近达到 CRLB（其渐近协方差矩阵为 $I(\theta)^{-1}$，其中 $I(\theta)$ 为估计 θ 的 Fisher 信息矩阵）.本章第二部分的主题是证明（为简化起见，仅讨论一维情况）$I(\theta)$ 确实是渐近有效方差.

你可能会问，为什么渐近有效方差与 $I(\theta)^{-1}$ 有关？可以这样理解：Fisher 信息通过研究映射

$$\theta \mapsto \log p_\theta$$

的导数得到.然而，真正起作用的是其逆映射

$$P \mapsto \theta$$

或，当参数 $\gamma = g(\theta)$ 为研究重点时，对应的映射

$$P \mapsto \gamma$$

回想一下函数的逆导数性质（例如 $f : \mathbb{R} \to \mathbb{R}$）：其导数的逆是逆函数的导数.在我们的情境中，映射 $P \mapsto \gamma$ 是相当抽象的，因此研究其导数需要引入一些新的概念，这也是本章第一部分的研究重点.

15.1 替代估计量 ⋆

当 $\mathcal{X}$ 是欧几里得空间时，可以定义分布函数 $F(x) := P_\theta(X \leq x)$ 以及经验分布函数

$$\hat{F}_n(x) = \frac{1}{n} \# \{X_i \leq x, 1 \leq i \leq n\}$$

这实际上是一个概率分布函数，其在每个观测值处分配 $1/n$ 的概率质量.对于一般的 $\mathcal{X}$，定义经验分布 $\hat{P}_n$ 为在每个观测值处分配质量 $1/n$ 的分布，更正式地表示为：

$$\hat{P}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

其中 δ_x 为 x 处的点质量.因此，对于 $\mathcal{X}$ 的（可测）子集 A：

$$\hat{P}_n(A) = \frac{1}{n} \# \{X_i \in A, 1 \leq i \leq n\}$$

对于（可测）函数 $f : \mathcal{X} \to \mathbb{R}^r$（$r \in \mathbb{N}$），我们定义：

$$\hat{P}_n f := \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \int f d\hat{P}_n$$

类似地，对于集合 A，有：

$$\hat{P}_n(A) = \hat{P}_n l_A$$

同样，使用 P_θ 下的期望符号表示：

$$P_\theta f := \mathbb{E}_\theta f(X) = \int f dP_\theta, \quad P_\theta(A) = P_\theta \mathrm{l}_A$$

感兴趣的参数 γ 通常表示为：

$$\gamma = g(\theta) \in \mathbb{R}^p$$

它往往可以写为形式：

$$\gamma = Q(P_\theta),$$

其中 Q 是模型类 $\mathcal{P}$ 上某个泛函.假设 Q 在经验测度 $\hat{P}_n$ 上也有定义，那么 γ 的替代估计量为：

$$T_n := Q(\hat{P}_n).$$

反之，

定义 15.1.1 如果统计量 T_n 可以表示为 $T_n = Q(\hat{P}_n)$，且对所有 $\theta \in \Theta$ 有 $Q(P_\theta) = g(\theta)$，则称其为 $\gamma = g(\theta)$ 的 Fisher 一致估计量.

我们也会遇到以下修正形式：

$$T_n = Q_n(\hat{P}_n)$$

且当 n 足够大时：

$$Q_n(P_\theta) \approx Q(P_\theta) = g(\theta)$$

例子 15.1.1 均值函数的替代估计量

设定给定函数 $f : \mathcal{X} \to \mathbb{R}^r$ 和 $h : \mathbb{R}^r \to \mathbb{R}^p$，定义 $\gamma := h(P_\theta f)$.对应的替代估计量为：

$$T_n = h(\hat{P}_n f).$$

例子 15.1.2 M-估计和 Z-估计是替代估计量

M-估计量：

$$\hat{\gamma}_n := \arg\min_{c \in \Gamma} \hat{P}_n \rho_c$$

是以下形式的替代估计量：

$$\gamma = \arg\min_{c\in\Gamma} P_\theta \rho_c.$$

类似地，Z-估计量 $\hat{\gamma}_n$ 为以下方程的解：

$$\hat{P}_n \psi_c \Big|_{c=\hat{\gamma}_n} = 0$$

是以下形式的替代估计量：

$$P_\theta \psi_c|_{c=\gamma} = 0$$

例子 15.1.3 α 截尾均值的替代估计量

令 $\mathcal{X} = \mathbb{R}$，α 截尾均值定义为：

$$T_n := \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)},$$

其中 $X_{(i)}$ 是第 i 个顺序统计量，可写为：

$$X_{(i)} = \hat{F}_n^{-1}(i/n),$$

进而写为：

$$T_n = \frac{1}{1-2\alpha_n} \int_{\alpha_n+1/n}^{1-\alpha_n} \hat{F}_n^{-1}(u)du := Q_n(\hat{P}_n),$$

其中 $\alpha_n = [n\alpha]/n$.将 $\hat{F}_n$ 替换为 F，得：

$$Q_n(F) \approx \frac{1}{1-2\alpha} \int_\alpha^{1-\alpha} F^{-1}(u)du := Q(P_\theta).$$

例子 15.1.4 直方图作为密度的替代估计量

设 $\mathcal{X} = \mathbb{R}$，假设 X 关于 Lebesgue 测度具有密度 f，且 f 为感兴趣的参数.密度 f 可写为：

$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h}$$

直接用 $\hat{F}_n$ 替代 F 是不可行的.因此，这种情况下 $Q(P) = f$ 仅对具有密度的分布 P 有定义.不过，可以稍作调整，使用以下估计量：

$$\hat{f}_n(x) := \frac{\hat{F}_n(x+h_n) - \hat{F}_n(x-h_n)}{2h_n} := Q_n(\hat{P}_n),$$

其中 h_n 较小，满足 $h_n \to 0$ 且 $n \to \infty$.

15.2 替代估计量的一致性

我们首先给出经验分布函数一致收敛于理论分布函数的结论.

类似的一致收敛结果在更广泛的情境下成立（见 14.2 证明 M-估计量一致性）.

定理 15.2.1 （Glivenko-Cantelli 定理）
令 $\mathcal{X}=\mathbb{R}$，则有：

$$\sup_x |\hat{F}_n(x)-F(x)| \to 0, \quad \mathbb{P}_\theta - \text{a.s.}$$

证明
根据大数定律，对于所有 x，有：

$$|\hat{F}_n(x)-F(x)| \to 0, \quad \mathbb{P}_\theta - \text{a.s.}$$

因此，对于有限点集 $a_1,\dots,a_N$，有：

$$\max_{1\le j\le N} |\hat{F}_n(a_j)-F(a_j)| \to 0, \quad \mathbb{P}_\theta - \text{a.s.}$$

设 $\epsilon>0$，选取 $a_0<a_1<\cdots<a_N$ 满足：

$$F(a_j)-F(a_{j-1}) \le \epsilon, \quad j=1,\dots,N,$$

其中 $F(a_0):=0$ 且 $F(a_N):=1$.当 $x\in(a_{j-1},a_j]$ 时，有：

$$\hat{F}_n(x)-F(x) \le \hat{F}_n(a_j)-F(a_{j-1}) \le \hat{F}_n(a_j)-F(a_j)+\epsilon,$$

以及

$$\hat{F}_n(x)-F(x) \ge \hat{F}_n(a_{j-1})-F(a_j) \ge \hat{F}_n(a_{j-1})-F(a_{j-1})-\epsilon.$$

因此：

$$\sup_x |\hat{F}_n(x)-F(x)| \le \max_{1\le j\le N} |\hat{F}_n(a_j)-F(a_j)|+\epsilon \to \epsilon, \quad \mathbb{P}_\theta - \text{a.s.}$$

例子 15.2.1 样本中位数的一致性

设 $\mathcal{X}=\mathbb{R}$，F 为 X 的分布函数，我们考虑估计中位数 $\gamma:=F^{-1}(1/2)$.假设 F 连续且严格单调递增.样本中位数定义为：

$$T_n:=\hat{F}_n^{-1}(1/2):=\begin{cases} X_{((n+1)/2)} & n \text{ 奇数}, \\ \left[X_{(n/2)}+X_{(n/2+1)}\right]/2 & n \text{ 偶数}. \end{cases}$$

因此，有：

$$\hat{F}_n(T_n) = \frac{1}{2} + \begin{cases} 1/(2n) & n \text{ 奇数}, \\ 0 & n \text{ 偶数}. \end{cases}$$

从而：

$$\begin{aligned} |F(T_n) - F(\gamma)| &\le |\hat{F}_n(T_n) - F(T_n)| + |\hat{F}_n(T_n) - F(\gamma)| \\ &= |\hat{F}_n(T_n) - F(T_n)| + |\hat{F}_n(T_n) - 1/2| \\ &\le |\hat{F}_n(T_n) - F(T_n)| + 1/(2n) \to 0, \quad \mathbb{P}_\theta - \text{a.s.} \end{aligned}$$

因此，$\hat{F}_n^{-1}(1/2) = T_n \to \gamma = F^{-1}(1/2), \mathbb{P}_\theta$ -a.s.，即样本中位数是一致估计量.

15.3 替代估计量的渐近正态性*

设 $\gamma := Q(P) \in \mathbb{R}^p$ 为感兴趣的参数.本节的核心是将 δ -方法推广到非参数框架中.参数 δ -方法表明，如果 $\hat{\theta}_n$ 是参数 $\theta \in \mathbb{R}^p$ 的渐近线性估计量，而 $\gamma = g(\theta)$ 为 θ 的某个函数，且 g 在 θ 处可微，则 $\hat{\gamma}$ 是 γ 的渐近线性估计量.

现在，我们将 $\gamma = Q(P)$ 表示为概率测度 P 的函数（$P = P_\theta$，因此 $g(\theta) = Q(P_\theta)$）.我们让 P 起到类似于 θ 的作用，即用概率测度本身作为 $\mathcal{P}$ 的参数化.这需要重新定义导数，在抽象设置下，我们相对于 P 进行微分.

定义 15.3.1

- **影响函数**
  在 P 处，Q 的影响函数定义为：

  $$l_P(x) := \lim_{\epsilon \downarrow 0} \frac{Q\left((1-\epsilon)P + \epsilon\delta_x\right) - Q(P)}{\epsilon}, \quad x \in \mathcal{X},$$

  如果该极限存在.
- **Gâteaux 可微性**
  若对所有概率测度 $\tilde{P}$，有：

  $$\lim_{\epsilon \downarrow 0} \frac{Q\left((1-\epsilon)P + \epsilon\tilde{P}\right) - Q(P)}{\epsilon} = \mathbb{E}_{\tilde{P}} l_P(X),$$

  则称 Q 在 P 处是 Gâteaux 可微的.
- **Fréchet 可微性**
  设 d 是概率测度空间上的某个（伪）度量.若对任意概率测度 $\tilde{P}$，满足：

  $$Q(\tilde{P}) - Q(P) = \mathbb{E}_{\tilde{P}} l_P(X) + o(d(\tilde{P}, P)),$$

  则称 Q 在度量 d 下在 P 处是 Fréchet 可微的.

注释

1. 按照之前引入的记号，对于函数 $f:\mathcal{X}\to\mathbb{R}^r$ 和概率测度 $\tilde{P}$，定义如下期望：

$$\tilde{P}f := \mathbb{E}_{\tilde{P}}f(X)$$

2. 如果 Q 在 P 处是 Fréchet 或 Gâteaux 可微的，则有：

$$Pl_P := \mathbb{E}_P l_P(X) = 0$$

3. 如果 Q 在 P 处是 Fréchet 可微的，且满足：

$$d((1-\epsilon)P+\epsilon\tilde{P},P)=o(\epsilon),\quad \epsilon\downarrow 0,$$

则 Q 在 P 处是 Gâteaux 可微的：

$$\begin{aligned}Q((1-\epsilon)P+\epsilon\tilde{P})-Q(P) &= ((1-\epsilon)P+\epsilon\tilde{P})l_P+o(\epsilon)\\ &= \epsilon\tilde{P}l_P+o(\epsilon)\end{aligned}$$

定理与推论

引理 15.3.1

若 Q 在 P 处是 Fréchet 可微的，影响函数为 l_P，且满足

$$d\left(\hat{P}_n,P\right)=\mathcal{O}_{\mathbf{P}}\left(n^{-1/2}\right),$$

则有：

$$Q\left(\hat{P}_n\right)-Q(P)=\hat{P}_n l_P+o_{\mathbf{P}}\left(n^{-1/2}\right)$$

证明

由 Fréchet 可微性的定义直接得出.

推论 15.3.1

在满足引理 15.3.1 的条件下，若影响函数 l_P 满足 $V_P := Pl_P l_P^T<\infty$，则：

$$\sqrt{n}\left(Q\left(\hat{P}_n\right)-Q(P)\right)\xrightarrow{\mathcal{D}_P}\mathcal{N}\left(0,V_P\right)$$

例子：条件 (15.1) 的成立性

设 $\mathcal{X}=\mathbb{R}$，定义度量：

$$d(\tilde{P},P):=\sup_x|\tilde{F}(x)-F(x)|$$

根据 Donsker 定理可得：

$$d\left(\hat{P}_n, P\right) = O_{\mathbf{P}}\left(n^{-1/2}\right).$$

Donsker **定理**

若分布函数 F 是连续的，则：

$$\sup_x \sqrt{n}\left|\hat{F}_n(x) - F(x)\right| \xrightarrow{\mathcal{D}} Z,$$

其中随机变量 Z 的分布函数为：

$$G(z) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j+1}\exp\left[-2j^2z^2\right], \quad z \geq 0.$$

例子 15.3.1 Z-估计量的渐近线性性

设 γ 满足：

$$P\psi_\gamma = 0$$

定义：

$$P_\epsilon := (1-\epsilon)P + \epsilon\tilde{P}$$

设 γ_ϵ 满足：

$$P_\epsilon\psi_{\gamma_\epsilon} = 0$$

假设 $\gamma_\epsilon \to \gamma$，当 $\epsilon \downarrow 0$ 时，有：

$$P\left(\psi_{\gamma_\epsilon} - \psi_\gamma\right) + \epsilon(\tilde{P} - P)\psi_{\gamma_\epsilon} = 0$$

假设 $c \mapsto P\psi_c$ 可微，则：

$$\begin{aligned} P\left(\psi_{\gamma_\epsilon} - \psi_\gamma\right) &= \left(\frac{\partial}{\partial c^T}P\psi_c\Big|_{c=\gamma}\right)(\gamma_\epsilon - \gamma) + o(|\gamma_\epsilon - \gamma|) \\ &:= M_P(\gamma_\epsilon - \gamma) + o(|\gamma_\epsilon - \gamma|) \end{aligned}$$

结合其他假设，可得：

$$(\gamma_\epsilon - \gamma)(1 + o(1)) = -\epsilon M_P^{-1}\tilde{P}\psi_\gamma + o(\epsilon)$$

故：

$$\frac{\gamma_\epsilon - \gamma}{\epsilon} \to -M_P^{-1}\tilde{P}\psi_\gamma$$

影响函数为：

$$l_P = -M_P^{-1}\psi_\gamma$$

例子 15.3.2 α -截尾均值的渐近线性性

α -截尾均值是以下参数的替代估计量：

$$\gamma := Q(P) = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x)$$

使用分部积分可得：

$$(1-2\alpha)\gamma = (1-\alpha)F^{-1}(1-\alpha) - \alpha F^{-1}(\alpha) - \int_{\alpha}^{1-\alpha} v dF^{-1}(v)$$

利用分位数 $F^{-1}(v)$ 的影响函数：

$$q_v(x) = -\frac{1}{f\left(F^{-1}(v)\right)}\left(1\{x \le F^{-1}(v)\} - v\right),$$

可推导出：

$$l_P(x) = -\frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} (1\{x \le u\} - F(u))du$$

因此，α -截尾均值在正则条件下是渐近线性的，影响函数为 l_P，渐近正态，方差为 Pl_P^2.

15.4 渐近 Cramér Rao 下界 ⋆

设 X 的分布为 $P \in \{P_\theta : \theta \in \Theta\}$.为简化假设，令 $\Theta \subset \mathbb{R}$，$\theta$ 为感兴趣的参数.令 T_n 为 θ 的估计量.

本节假设一定的正则性条件（部分未具体说明）.尤其是，我们假设 $\mathcal{P}$ 由某个 σ -有限测度 ν 主导，并且 Fisher 信息量

$$I(\theta) := \mathbb{E}_\theta s_\theta^2(X)$$

对所有 θ 存在.其中，s_θ 为得分函数：

$$s_\theta := \frac{d}{d\theta}\log p_\theta = \frac{\dot{p}_\theta}{p_\theta}$$

其中 $p_\theta := \frac{dP_\theta}{d\nu}$.

回忆，如果 T_n 是 θ 的无偏估计量，则根据 Cramér Rao 下界，$1/I(\theta)$ 是其方差的下界（见 5.5 节的正则性条件 I 和 II）.

定义 15.4.1 渐近偏差与渐近方差

如果

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(b_\theta, V_\theta), \quad \forall\theta,$$

则称 b_θ 为渐近偏差，V_θ 为渐近方差.

- 若 $b_\theta = 0, \forall \theta$，则称 T_n 渐近无偏.
- 若 T_n 渐近无偏，且 $V_\theta = 1/I(\theta), \forall \theta$ 且某些正则性条件成立，则称 T_n 渐近有效.

注释

1. **避免超效率**
   上述定义假设条件对所有 θ 成立.对于固定的 θ_0，可以轻松构造“超效率”估计量，例如 $T_n = \theta_0$.因此，需要条件在所有 θ 上或所有序列 $\{\theta_n\}$ 上成立，甚至允许 $\theta_n = \theta + h/\sqrt{n}$.具体数学描述见 van der Vaart (1998)，略见于 Le Cam 第三引理.
2. **变动的 θ_n**
   若 $\theta = \theta_n$ 随 n 变化，则 X_i 的分布随 n 变化.因此，可以将样本 $X_1, \ldots, X_n$ 视为每个 n 的新样本 $X_{1,1}, \ldots, X_{n,n}$.
3. **MLE 的渐近有效性**
   通常，MLE $\hat{\theta}_n$ 渐近无偏，且渐近方差为 $1/I(\theta)$，即在正则性条件下，MLE 渐近有效.

对于具有影响函数 l_θ 的渐近线性估计量，其渐近方差为 $V_\theta = \mathbb{E}_\theta l_\theta^2(X)$.下一引理表明，$1/I(\theta)$ 是渐近方差的下界.

引理 15.4.1 渐近方差下界

若 T_n 满足渐近线性性：

$$T_n - \theta = \frac{1}{n} \sum_{i=1}^{n} l_\theta(X_i) + o_{\mathbf{P}_\theta}(n^{-1/2}),$$

且

$$\mathbb{E}_\theta l_\theta(X) = 0, \quad \mathbb{E}_\theta l_\theta^2(X) := V_\theta < \infty,$$

并且

$$\mathbb{E}_\theta l_\theta(X) s_\theta(X) = 1, \tag{15.2}$$

则

$$V_\theta \geq \frac{1}{I(\theta)}.$$

证明

由 Cauchy-Schwarz 不等式得：

$$\begin{aligned} 1 &= \left| \operatorname{cov}_\theta(l_\theta(X), s_\theta(X)) \right|^2 \\ &\leq \operatorname{var}_\theta(l_\theta(X)) \operatorname{var}_\theta(s_\theta(X)) = V_\theta I(\theta). \end{aligned}$$

例子 15.4.1 Z-估计量的等式 (15.2)

对于 Z-估计量，θ 满足：

$$P\psi_\theta = 0$$

其影响函数为：

$$l_\theta = -\frac{\psi_\theta}{M_\theta}$$

其中

$$M_\theta := \frac{d}{d\theta} P\psi_\theta$$

在正则性条件下，

$$M_\theta = P\dot{\psi}_\theta = \int \dot{\psi}_\theta p_\theta d\nu, \quad \dot{\psi}_\theta := \frac{d}{d\theta}\psi_\theta.$$

通过链式法则，也可写为：

$$M_\theta = -\int \psi_\theta \dot{p}_\theta d\nu, \quad \dot{p}_\theta := \frac{d}{d\theta} p_\theta.$$

因此，有：

$$Pl_\theta s_\theta = -M_\theta^{-1} P\psi_\theta s_\theta = -M_\theta^{-1} \int \psi_\theta \dot{p}_\theta d\nu = 1,$$

即等式 (15.2) 成立.

例子 15.4.2 "插入"估计量的等式 (15.2)

我们现在考虑"插入"估计量 $Q\left(\hat{P}_n\right)$.假设 Q 是 Fisher 一致的（即，对于所有 θ，$Q\left(P_\theta\right) = \theta$）. 进一步假设 Q 在所有 P_θ 处相对于度量 d 是 Fréchet 可微的，并且满足：

$$d\left(P_{\tilde{\theta}}, P_\theta\right) = \mathcal{O}(|\tilde{\theta} - \theta|).$$

根据 Fréchet 可微性的定义，

$$h = Q\left(P_{\theta+h}\right) - Q\left(P_\theta\right) = P_{\theta+h} l_\theta + o(|h|) = (P_{\theta+h} - P_\theta) l_\theta + o(|h|),$$

或者，当 $h \to 0$ 时，

$$\begin{aligned} 1 &= \frac{(P_{\theta+h} - P_\theta) l_\theta}{h} + o(1) = \frac{\int l_\theta \left(p_{\theta+h} - p_\theta\right) d\nu}{h} + o(1) \\ &\to \int l_\theta \dot{p}_\theta d\nu = P_\theta\left(l_\theta s_\theta\right). \end{aligned}$$

因此，(15.2) 成立.

15.5 Le Cam 第三引理 ⋆

以下例子旨在说明，点态渐近性质可能导致误导，因此需要考虑依赖于样本量 n 的参数序列 θ_n.

例子 15.5.1 Hodges-Lehmann 的超效率例子

设 $X_1, \dots, X_n$ 是 X 的独立同分布样本，其中 $X = \theta + \epsilon$，且 $\epsilon \sim \mathcal{N}(0,1)$.考虑估计量：

$$T_n := \begin{cases} \bar{X}_n, & \text{若 } |\bar{X}_n| > n^{-1/4}, \\ \bar{X}_n/2, & \text{若 } |\bar{X}_n| \le n^{-1/4}. \end{cases}$$

则有：

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}_\theta} \begin{cases} \mathcal{N}(0,1), & \theta \neq 0, \\ \mathcal{N}(0,1/4), & \theta = 0. \end{cases}$$

因此，点态渐近分析显示 T_n 在 $\theta = 0$ 时比样本均值 $\bar{X}_n$ 更有效.但如果考虑参数序列 θ_n（例如 $\theta_n = h/\sqrt{n}$）会发生什么?

在 $\mathbb{P}_{\theta_n}$ 下，$\bar{X}_n = \bar{\epsilon}_n + h/\sqrt{n} = \mathcal{O}_{\mathbb{P}_{\theta_n}}(n^{-1/2})$，从而 $\mathbb{P}_{\theta_n}(|\bar{X}_n| > n^{-1/4}) \to 0$，即 $\mathbb{P}_{\theta_n}(T_n = \bar{X}_n) \to 0$.

因此，

$$\begin{aligned} \sqrt{n}(T_n - \theta_n) &= \sqrt{n}(T_n - \theta_n)1\{T_n = \bar{X}_n\} \\ &\quad + \sqrt{n}(T_n - \theta_n)1\{T_n = \bar{X}_n/2\} \\ &\xrightarrow{\mathcal{D}_{\theta_n}} \mathcal{N}\left(-\frac{h}{2}, \frac{1}{4}\right). \end{aligned}$$

渐近均方误差（AMSE）定义为渐近方差加渐近偏差的平方：

$$\mathrm{AMSE}_{\theta_n}(T_n) = \frac{1 + h^2}{4}.$$

样本均值 $\bar{X}_n$ 的 AMSE 是其标准化非渐近均方误差：

$$\mathrm{AMSE}_{\theta_n}(\bar{X}_n) = \mathrm{MSE}_{\theta_n}(\bar{X}_n) = 1.$$

因此，当 h 足够大时，T_n 的渐近均方误差比 $\bar{X}_n$ 更大.

Le Cam 第三引理

Le Cam 的第三引理表明，对所有 θ 的渐近线性性意味着渐近正态性，也适用于参数序列 $\theta_n = \theta + h/\sqrt{n}$.对于这样的序列，渐近方差不变.此外，如果 (15.2) 对所有 θ 成立，则估计量在 $\mathbb{P}_{\theta_n}$ 下渐近无偏.

引理 15.5.1 Le Cam 第三引理

假设对所有 θ，有：

$$T_n - \theta = \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_i) + o_{\mathbf{P}_\theta}(n^{-1/2}),$$

其中 $P_\theta l_\theta = 0$, 且 $V_\theta := P_\theta l_\theta^2 < \infty$.则在正则性条件下，

$$\sqrt{n}(T_n - \theta_n) \xrightarrow{\mathcal{D}_{\theta_n}} \mathcal{N}\left(\{P_\theta(l_\theta s_\theta) - 1\}h, V_\theta\right).$$

我们将简要证明此引理.为此需要以下辅助引理：

引理 15.5.2 辅助引理

设 $Z \in \mathbb{R}^2$ 服从 $\mathcal{N}(\mu, \Sigma)$，其中

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix},$$

且满足 $\mu_2 = -\sigma_2^2/2$.

设 $Y \in \mathbb{R}^2$ 服从 $\mathcal{N}(\mu + a, \Sigma)$，其中

$$a = \begin{pmatrix} \sigma_{1,2} \\ \sigma_2^2 \end{pmatrix}.$$

设 Z 的密度为 ϕ_Z，Y 的密度为 ϕ_Y.则对任意 $z = (z_1, z_2) \in \mathbb{R}^2$，有：

$$\phi_Z(z) e^{z_2} = \phi_Y(z).$$

证明

Z 的密度函数为：

$$\phi_Z(z) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right].$$

定义 $a = \begin{pmatrix} \sigma_{1,2} \\ \sigma_2^2 \end{pmatrix}$，计算 $\Sigma^{-1}a$，可以得到：

$$\Sigma^{-1}a = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

从而，我们对二次型分解，得到：

$$\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu) = \frac{1}{2}(z-\mu-a)^T\Sigma^{-1}(z-\mu-a) + a^T\Sigma^{-1}(z-\mu) - \frac{1}{2}a^T\Sigma^{-1}a.$$

接着计算偏差项 $a^T\Sigma^{-1}(z-\mu)$：

$$a^T\Sigma^{-1}(z-\mu)-\frac{1}{2}a^T\Sigma^{-1}a=(0\quad 1)(z-\mu)-\frac{1}{2}(0\quad 1)a.$$

由于 $\mu_2=-\sigma_2^2/2$，因此有：

$$z_2-\mu_2-\frac{1}{2}\sigma_2^2=z_2.$$

代入密度函数中，我们发现：

$$\phi_Z(z)\exp(z_2)=\phi_Y(z).$$

结论

引理证明完成，Z 的密度与 Y 的密度通过 $\exp(z_2)$ 相互关联.这个结果将在后续推导 Le Cam 第三引理时用到.

Le Cam 第三引理证明概要

设

$$\Lambda_n:=\sum_{i=1}^n[\log p_{\theta_n}(X_i)-\log p_\theta(X_i)]$$

在 $\mathbb{P}_\theta$ 下，通过二阶泰勒展开，有：

$$\begin{aligned}\Lambda_n&\approx\frac{h}{\sqrt{n}}\sum_{i=1}^n s_\theta(X_i)+\frac{h^2}{2}\frac{1}{n}\sum_{i=1}^n\dot{s}_\theta(X_i)\\&\approx\frac{h}{\sqrt{n}}\sum_{i=1}^n s_\theta(X_i)-\frac{h^2}{2}I(\theta),\end{aligned}$$

其中利用了

$$\frac{1}{n}\sum_{i=1}^n\dot{s}_\theta(X_i)\approx\mathbb{E}_\theta\dot{s}_\theta(X)=-I(\theta)$$

此外，根据假设的渐近线性性，在 $\mathbb{P}_\theta$ 下，

$$\sqrt{n}(T_n-\theta)\approx\frac{1}{\sqrt{n}}\sum_{i=1}^n l_\theta(X_i)$$

因此，

$$\begin{pmatrix}\sqrt{n}(T_n-\theta)\\\Lambda_n\end{pmatrix}\xrightarrow{\mathcal{D}_\theta}Z,$$

其中 $Z\in\mathbb{R}^2$ 满足二元正态分布：

$$Z=\begin{pmatrix}Z_1\\Z_2\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}0\\-\frac{h^2}{2}I(\theta)\end{pmatrix},\begin{pmatrix}V_\theta&hP_\theta(l_\theta s_\theta)\\hP_\theta(l_\theta s_\theta)&h^2I(\theta)\end{pmatrix}\right)$$

由此，对于任意有界连续函数 $f:\mathbb{R}^2\to\mathbb{R}$，有：

$$\mathbb{E}_\theta f\left(\sqrt{n}(T_n-\theta),\Lambda_n\right)\to\mathbb{E}f(Z_1,Z_2)$$

设 $f:\mathbb{R}\to\mathbb{R}$ 为有界连续函数，由于

$$\prod_{i=1}^n p_{\theta_n}(X_i)=\prod_{i=1}^n p_\theta(X_i)e^{\Lambda_n},$$

可以写为：

$$\mathbb{E}_{\theta_n} f\left(\sqrt{n}(T_n-\theta)\right)=\mathbb{E}_\theta f\left(\sqrt{n}(T_n-\theta)\right)e^{\Lambda_n}$$

应用扩展的 Portmanteau 定理，可以得到：

$$\mathbb{E}_\theta f\left(\sqrt{n}(T_n-\theta)\right)e^{\Lambda_n}\to\mathbb{E}f(Z_1)e^{Z_2}$$

结合辅助引理，其中

$$\mu=\begin{pmatrix}0\\-\frac{h^2}{2}I(\theta)\end{pmatrix},\quad \Sigma=\begin{pmatrix}V_\theta & hP_\theta(l_\theta s_\theta)\\ hP_\theta(l_\theta s_\theta) & h^2I(\theta)\end{pmatrix}$$

我们有：

$$\mathbb{E}f(Z_1)e^{Z_2}=\int f(z_1)e^{z_2}\phi_Z(z)dz=\int f(z_1)\phi_Y(z)dz=\mathbb{E}f(Y_1),$$

其中

$$Y=\begin{pmatrix}Y_1\\Y_2\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}hP_\theta(l_\theta s_\theta)\\ \frac{h^2}{2}I(\theta)\end{pmatrix},\begin{pmatrix}V_\theta & hP_\theta(l_\theta s_\theta)\\ hP_\theta(l_\theta s_\theta) & h^2I(\theta)\end{pmatrix}\right)$$

所以，

$$Y_1\sim\mathcal{N}\left(hP_\theta(l_\theta s_\theta),V_\theta\right)$$

最终得到结论：

$$\sqrt{n}(T_n-\theta)\xrightarrow{\mathcal{D}_{\theta_n}}Y_1\sim\mathcal{N}\left(hP_\theta(l_\theta s_\theta),V_\theta\right)$$

因此，

$$\sqrt{n}(T_n-\theta_n)=\sqrt{n}(T_n-\theta)-h\xrightarrow{\mathcal{D}_{\theta_n}}\mathcal{N}\left(h\{P_\theta(l_\theta s_\theta)-1\},V_\theta\right)$$

第 16 章：复杂度正则化

在我们假设有 $X_1,\ldots,X_n$ 作为独立同分布样本，随机变量 X 的分布 P 被建模为 $P\in\mathcal{P}=\{P_\theta:\theta\in\Theta\}$.若 Θ 是有限维的，其维度 p 可以被看作参数空间 Θ 复杂度的度量.如果

$p > n$，则参数比观测值更多，这种情况直观上就不是统计上理想的情况.这与未知数多于方程的系统相似，属于病态系统.

拥有过多参数的模型可能会很好地拟合数据，但预测能力却可能很差，容易出现过拟合.复杂度正则化是应对复杂参数空间的一种方法，让数据决定哪个子模型在逼近误差和估计误差之间提供了良好的权衡.

即使参数空间是 ∞ 维的，也不总是需要应用复杂度正则化.参数空间 Θ 的维度本身并非总是复杂度的最佳描述.如果 Θ 是度量空间，可以应用所谓的熵来测量复杂度.由于篇幅限制，这些细节不在讲义范围内.我们接下来以非参数回归和高维回归问题为原型例子.

16.1 非参数回归

考虑 n 个实值响应变量 $Y_1, \ldots, Y_n$，它们依赖于一些固定的协变量 $x_1, \ldots, x_n$，这些协变量对所有 i 均在某个空间 $\mathcal{X}$ 内，形式为：

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

其中 $\epsilon_1, \ldots, \epsilon_n$ 是不可观测的噪声，f 是未知函数.

假设我们尝试使用最小二乘估计量来估计 f.令 $Y := (Y_1, \ldots, Y_n)^T$.为了方便，我们将函数 $f : \mathcal{X} \to \mathbb{R}$ 表示为向量：

$$\mathrm{f} := (f(x_1), \ldots, f(x_n))^T \in \mathbb{R}^n$$

最小二乘估计量定义为：

$$\hat{f}_{\mathrm{LS}} := \arg\min_{\mathrm{f} \in \mathbb{R}^n} \|Y - \mathrm{f}\|_2^2$$

显然，如果所有 x_i 都不同，则 $\hat{f}_{\mathrm{LS}} = Y$，并且存在完全拟合：

$$\|Y - \hat{f}_{\mathrm{LS}}\|_2^2 = 0$$

这是过拟合的典型实例.估计量 $\hat{f}_{\mathrm{LS}}$ 仅仅复现了数据，完全没有预测能力.

因此，我们需要为 f 提供一个模型，假设 $f \in \mathcal{F}$，其中 $\mathcal{F}$ 是某个函数类.

16.2 光滑类

假设 $\mathcal{X}$ 是 $\mathbb{R}$ 中的某个区间，例如 $\mathcal{X} = [0, 1]$.根据具体情况，可以合理地假设函数 f 不会过于“起伏”.一种数学描述方法是假设 f 可微，且其导数 f' 的绝对值在一定范围内.例如，可以通过 f 的导数的 (Sobolev) 半范数来衡量其粗糙程度：

$$\sqrt{\int_0^1 |f'(x)|^2 dx}$$

一种方法是对所有满足 $\int_0^1 |f'(x)|^2 dx \le M^2$ 的 f 进行最小二乘拟合，其中 M 是给定常数.更灵活的方法是使用拉格朗日版本.选取一个调节参数 $\lambda \ge 0$，定义估计量 $\hat{f}$ 为

$$\hat{f} := \arg\min_f \left\{ \|Y - f\|_2^2 + \lambda^2 \int_0^1 |f'(x)|^2 dx \right\}$$

这种方法称为 Tikhonov 正则化的一种形式.其中，

$$\lambda^2 \int_0^1 |f'(x)|^2 dx$$

被视为选择过于“起伏”函数的惩罚项.这一惩罚项正则化了函数，调节参数 λ 控制正则化的程度：λ 越大，估计量越平滑.选择 λ 是一个难点.从理论上来看，有一些指导意见（例如，选择 λ^2 的阶为 $n^{1/3}$，以平衡逼近误差和估计误差）.也可以使用贝叶斯方法选择 λ.在实践中，可以采用交叉验证.

目前，我们将光滑性描述为一阶导数有界.如果我们认为未知函数 f 存在更高阶导数，也可以使用更高阶导数.设 $f^{(m)}$ 表示函数 $f : [0,1] \to \mathbb{R}$ 的 m 阶导数.那么可能的惩罚项为：

$$\lambda^2 \int_0^1 |f^{(m)}(x)|^2 dx$$

对于较高阶的 m，可以选择较小的 λ（例如，选择 $\lambda \sim n^{-1/(2m+1)}$，以平衡逼近误差和估计误差）.所得估计量称为平滑样条.

对于二次惩罚，带惩罚的最小二乘估计量 $\hat{f}$ 的计算并不困难，因为它是二次函数的极小值.然而，通常无法获得显式表达式.在下一节中，我们将提供连续版本的显式解，作为“趣味”.

16.3 显式解的连续版本 ⋆

我们研究上一节的连续版本.此问题可以显式求解.假设我们观察到一个函数 $y : [0,1] \to \mathbb{R}$，并希望通过上一节的惩罚方法对其进行平滑.令

$$\hat{f} = \arg\min_f \left\{ \int_0^1 |y(x) - f(x)|^2 dx + \lambda^2 \int_0^1 |f'(x)|^2 dx \right\} \tag{16.1}$$

引理 16.3.1 令 $\hat{f}$ 满足 (16.1).则有：

$$\hat{f}(x) = \frac{C}{\lambda} \cosh\left(\frac{x}{\lambda}\right) + \frac{1}{\lambda} \int_0^x y(u) \sinh\left(\frac{u-x}{\lambda}\right) du$$

其中，

$$C = Y(1) - \left\{ \frac{1}{\lambda} \int_0^1 Y(u) \sinh\left(\frac{1-u}{\lambda}\right) du \right\} / \sinh\left(\frac{1}{\lambda}\right)$$

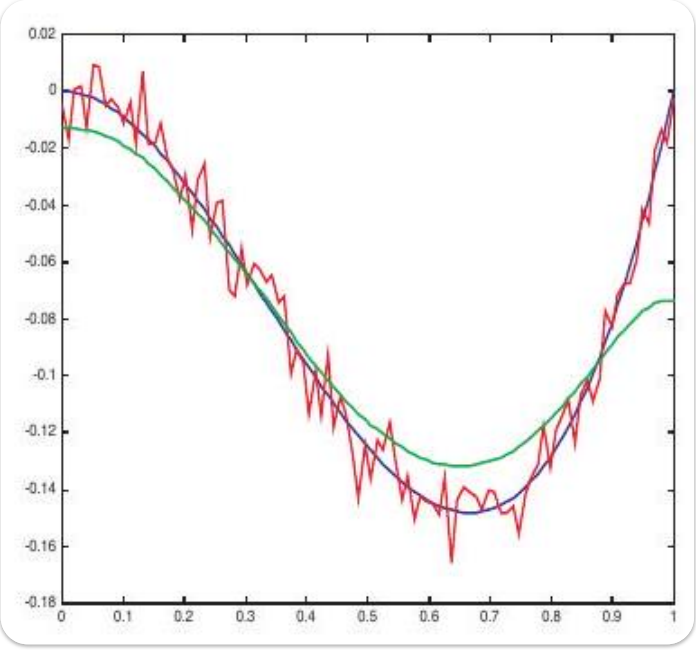
并且

$$Y(x) = \int_0^x y(u) du$$

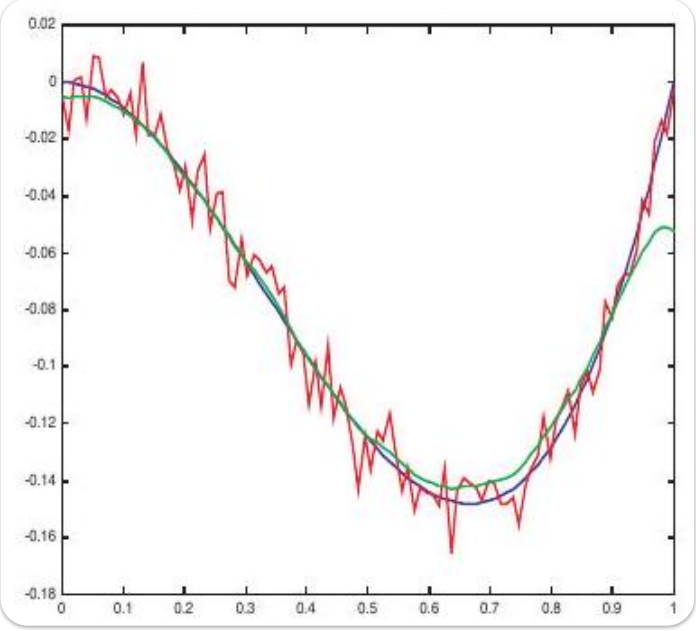
此证明基于变分法，此处略去.

借助这一显式表达式，可以轻松实现估计过程.以下是一个例子：

数值例子



去噪后，$\lambda = 0.1$，误差 $= 2.8119 \times 10^{-4}$



黑色曲线为未知函数 f（由于这是模拟，我们知道 f）.红色波动曲线为观察到的函数 y，绿色曲线为估计量 $\hat{f}$.从第二张图可以看出，通过减小调节参数 λ，估计量 $\hat{f}$ 更接近于未知的真实函数 f.

去噪后，$\lambda = 0.05$，误差 $= 7.8683 \times 10^{-5}$

16.4 有界变差函数的估计

假设仍然有 $\mathcal{X}=[0,1]$.除了采用惩罚项

$$\lambda^2\int_0^1|f'(x)|^2dx$$

外，还可以选择

$$\lambda^2\int_0^1|f'(x)|dx$$

即不加平方.这看似是一个微小的修改，但实际上差异非常大.进一步放宽可微性假设，定义函数 f 的总变差为：

$$\mathrm{TV}(f):=\sum_{i=2}^n|f(x_i)-f(x_{i-1})|$$

其中假设 $x_1<x_2<\ldots<x_n$.这引出了估计量：

$$\hat{f}:=\arg\min_f\left\{\|Y-f\|_2^2+\lambda^2\mathrm{TV}(f)\right\}$$

类比而言，如果 Y 表示某些地段的山高，而我们需要修一条路，则通过总变差惩罚可以平整山丘和山谷（使道路不会有陡坡），同时尽量减少挖土量（以最小二乘损失衡量）.这一过程可以通过迭代实现，例如在坡度最大的地方填补山谷或削平山峰.

这一估计量具有局部适应性：通过增大 λ，仅进行局部的改动.与第 16.2 节的估计量或第 16.3 节的连续版本不同，改变 λ 会对全局产生影响.

此外，可以将该估计量表述为具有 ℓ_1 惩罚的线性最小二乘问题的解.这一联系将有助于理解为什么去掉惩罚项中的平方会显著改变估计量的性质.

16.5 平滑与 Lasso 惩罚

定义 f 的形式化表达

定义 $\mathrm{f}(x_0):=0$.对于 $i=1,\ldots,n$，有：

$$\begin{aligned}\mathrm{f}(x_i)&=\sum_{j=1}^i\big(\mathrm{f}(x_j)-\mathrm{f}(x_{j-1})\big)\\&=\sum_{j=1}^n\underbrace{\big(\mathrm{f}(x_j)-\mathrm{f}(x_{j-1})\big)}_{:=b_j}\underbrace{1\{j\le i\}}_{:=\xi_{i,j}}\\&=\sum_{j=1}^n b_j\xi_{i,j}.\end{aligned}$$

将系数 $b_j(j=1,\ldots,n)$ 置于向量 $b=(b_1,\ldots,b_n)^T$ 中，并将 $\xi_{i,j}$ 放入矩阵

$$X := \begin{pmatrix} \xi_{1,1} & \cdots & \xi_{1,n} \\ \vdots & \ddots & \vdots \\ \xi_{n,1} & \cdots & \xi_{n,n} \end{pmatrix},$$

此时，带总变差惩罚的估计量可以写作：

$$\hat{f} = \arg\min_{\mathbf{f}=Xb} \left\{ \|Y - \mathbf{f}\|_2^2 + \lambda^2 \|b\|_1 \right\},$$

其中 $\|b\|_1 = \sum_{j=1}^{n} |b_j|$ 表示向量 $b \in \mathbb{R}^n$ 的 ℓ_1 范数.上述公式表明参数的数量等于观测数量，即 n.惩罚项保证即使在这种情况下，也不会过拟合数据（当 λ 不过小时）.

二维情况

接下来考虑 $\mathcal{X}$ 是二维情况，例如 $\mathcal{X} = [0,1]^2$.此时，未知函数 f 可以看作一幅图像，观测值为

$$Y_{i_1,i_2} = f(x_{i_1,i_2}) + \epsilon_{i_1,i_2},$$

假设有 $n = m^2$ 个观测点，它们分布在规则网格上：

$$x_{i_1,i_2} = \left(\frac{i_1}{m}, \frac{i_2}{m} \right), \quad i_1 = 1, \ldots, m,\ i_2 = 1, \ldots, m.$$

如何在图像上建模“平滑性”？可以采用类似于一维情况下的导数平方和惩罚 $\int |f'(x)|^2 dx$ 的形式.然而，与一维情况一样，平方项的惩罚具有非局部影响.基于导数平方和惩罚的图像重建可能会模糊图像.例如，对于包含湖泊和河流的景观图像，基于导数平方和的惩罚会使原本清晰的边界变得模糊.另一种替代方法是再次使用总变差惩罚.

在维度大于 1 的情况下，总变差有多种定义.对于二维情形，可定义为：

$$\mathrm{TV}(\mathbf{f}) := \sum_{i_1=2}^{m} \sum_{i_2=2}^{m} \left| \Delta \mathbf{f}(x_{i_1,i_2}) \right|,$$

其中

$$\Delta \mathbf{f}(x_{i_1,i_2}) := f\left(\frac{i_1}{m}, \frac{i_2}{m} \right) - f\left(\frac{i_1 - 1}{m}, \frac{i_2}{m} \right) - f\left(\frac{i_1}{m}, \frac{i_2 - 1}{m} \right) + f\left(\frac{i_1 - 1}{m}, \frac{i_2 - 1}{m} \right).$$

图像重建算法为：

$$\hat{f} := \arg\min_{\mathbf{f}} \left\{ \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \left(Y_{i_1,i_2} - f(x_{i_1,i_2}) \right)^2 + \lambda^2 \mathrm{TV}(f) \right\}.$$

该估计量同样可以写作带有 ℓ_1 惩罚项的线性函数最小二乘估计量.

线性模型中的 Ridge 和 Lasso 惩罚

对于线性模型，观测数据为 $(x_1, Y_1), \ldots, (x_n, Y_n)$，其中 $x_i \in \mathbb{R}^p$ 是 p 维行向量，$Y_i \in \mathbb{R}\,(i = 1, \ldots, n)$.目标是用最小二乘损失函数找到一个好的线性逼近：

$$b \mapsto \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{i,j} b_j \right)^2.$$

定义设计矩阵为：

$$X := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix},$$

响应向量为：

$$Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

设系数向量为 $b := (b_1 \quad \ldots \quad b_p)^T$，则损失函数为：

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{i,j} b_j \right)^2 = \|Y - Xb\|_2^2.$$

如果 $p \geq n$ 且 X 的秩为 n，则对所有 $b \in \mathbb{R}^p$ 最小化上述函数，会得到一个“完美”解 $\hat{\beta}_{\mathrm{LS}}$，使得 $X\hat{\beta}_{\mathrm{LS}} = Y$.但这种解仅仅复现了数据，因而毫无用处，称为“过拟合”.

定义 16.5.1 Ridge 回归估计量为：

$$\hat{\beta}_{\mathrm{ridge}} := \arg\min_{b \in \mathbb{R}} \left\{ \|Y - Xb\|_2^2 + \lambda^2 \|b\|_2^2 \right\},$$

其中 $\lambda > 0$ 是正则化参数.

定义 16.5.2 Lasso 估计量为：

$$\hat{\beta}_{\mathrm{Lasso}} := \arg\min_{b \in \mathbb{R}} \left\{ \|Y - Xb\|_2^2 + 2\lambda \|b\|_1 \right\},$$

其中 $\lambda > 0$ 是正则化参数，$\|b\|_1 := \sum_{j=1}^{p} |b_j|$ 是 b 的 ℓ_1 范数.

贝叶斯 MAP 估计的联系

回顾 10.5.3 节中的贝叶斯最大后验 (MAP) 估计定义.考虑模型 $Y = X\beta + \epsilon$，其中 $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.Ridge 回归估计量是先验分布为 $\beta_j \sim \mathcal{N}(0, \tau^2)$ 的 MAP 估计，其中 $\tau = \sigma/\lambda$ .Lasso 估计量是先验分布为 $\beta_j \sim \mathrm{Laplace}(0, \tau^2)$ 的 MAP 估计，其中标准差 $\tau = \sqrt{2}\sigma^2/\lambda$.详见 10.6 节.

备注

1. **关于收缩性**：随着正则化参数 λ 的增大，Ridge 回归会收缩回归系数，但这些系数不会精确地收缩到 0.而 Lasso 不仅会收缩系数，还会将部分（甚至多数）系数直接设为 0.当变量数目 p 较大时，通常优先使用 Lasso，因为信号低于噪声水平的变量更应被直接归零.
2. **偏差与方差**：Ridge 和 Lasso 的估计量都是有偏的.随着 λ 增大，偏差也会随之增大，但估计的方差会减小.
3. **正则化参数 λ 的选择**：通常可以通过交叉验证、信息论准则或贝叶斯方法选择 λ.理论上，对于 Lasso，一个合理的选择是 $\lambda \sim \sqrt{n \log p}$.

定理与推导

Ridge 回归估计量的推导

引理 16.5.1

Ridge 回归估计量为：

$$\hat{\beta}_{\text{ridge}} = \left(X^T X + \lambda^2 I\right)^{-1} X^T Y.$$

证明

目标函数为：

$$\|Y - Xb\|_2^2 + \lambda^2 \|b\|_2^2.$$

对 b 求导：

$$\frac{1}{2}\frac{\partial}{\partial b}\left\{\|Y - Xb\|_2^2 + \lambda^2 \|b\|_2^2\right\} = -X^T(Y - Xb) + \lambda^2 b = -X^T Y + (X^T X + \lambda^2 I)b.$$

令导数为 0，得：

$$\hat{\beta}_{\text{ridge}} = \left(X^T X + \lambda^2 I\right)^{-1} X^T Y.$$

正交设计下的特殊情况

推论 16.5.1

若 $X^T X = nI$（意味着 $p \leq n$），则：

$$\hat{\beta}_{\text{ridge}} = \frac{X^T Y}{n + \lambda^2}.$$

进一步分析，当 $\epsilon_1, \ldots, \epsilon_n$ 独立同分布，均值为 0，方差为 σ^2 时：

$$\mathbb{E}\|X\hat{\beta}_{\text{ridge}} - f\|_2^2 = \underbrace{\left[\frac{\lambda^2/n}{1 + \lambda^2/n}\right]^2 \|X\beta^*\|_2^2}_{\text{偏差}} + \underbrace{\left[\frac{1}{1 + \lambda^2/n}\right]^2 p\sigma^2}_{\text{方差}} + \underbrace{\|X\beta^* - f\|_2^2}_{\text{模型误差}},$$

其中，$X\beta^*$ 是 f 在 X 列向量张成空间上的投影.要平衡偏差和方差，需要已知噪声方差 σ^2 和 $\|X\beta^*\|_2^2$.然而，后者未知，因为 f 是未知的，这类似于 5.2.1 节中遇到的问题.

Lasso 估计量的推导

一般情况下，Lasso 估计量没有简单的解析表达式.以下为正交设计下的结果.

引理 16.5.2
若 $X^TX = nI$，令 $Z := X^TY$，则对 $j = 1, \dots, p$：

$$\hat{\beta}_{\text{Lasso},j} = \begin{cases} Z_j/n - \lambda/n, & Z_j \geq \lambda, \\ 0, & |Z_j| \leq \lambda, \\ Z_j/n + \lambda/n, & Z_j \leq -\lambda. \end{cases}$$

证明
简记 $\hat{\beta}_{\text{Lasso}} =: \hat{\beta}$，目标函数可写为：

$$\|Y - Xb\|_2^2 = \|Y\|_2^2 - 2b^TX^TY + nb^Tb = -2b^TZ + nb^Tb.$$

因此，对于每个 j，需最小化：

$$-2b_jZ_j + nb_j^2 + 2\lambda|b_j|.$$

若 $\hat{\beta}_j > 0$，则目标函数对 b_j 的导数为 0，即：

$$-Z_j + n\hat{\beta}_j + \lambda = 0,$$

解得：

$$\hat{\beta}_j = Z_j/n - \lambda/n.$$

类似地，若 $\hat{\beta}_j < 0$，有：

$$-Z_j + n\hat{\beta}_j - \lambda = 0,$$

解得：

$$\hat{\beta}_j = Z_j/n + \lambda/n.$$

若 $\hat{\beta}_j = 0$，需满足 $|Z_j| \leq \lambda$.

一些符号说明

1. 对于向量 $z \in \mathbb{R}^p$，其 ℓ_∞ 范数为：

$$\|z\|_\infty := \max_{1 \leq j \leq p} |z_j|.$$

2. 用 $X_1, \ldots, X_p$ 表示 X 的列向量.
3. 对于子集 $S \subset \{1, \ldots, p\}$，$X\beta_S^*$ 表示 $f = E[Y]$ 使用 S 中变量的最佳线性近似，即 $X\beta_S^*$ 是 f 在空间 $\{\sum_{j \in S} X_j b_{S,j} : b_S \in \mathbb{R}^{|S|}\}$ 上的投影.

定理 16.5.1

假设 $X^T X = nI$.设 $f = E[Y]$，$\epsilon = Y - f$，取某置信水平 $\alpha \in (0,1)$.若存在 λ_α，使得：

$$\mathbb{P}\left(\|X^T \epsilon\|_\infty > \lambda_\alpha\right) \leq \alpha,$$

则对于 $\lambda > \lambda_\alpha$，以至少 $1-\alpha$ 的概率，有：

$$\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2 \leq \min_S \left\{ \underbrace{\frac{(\lambda + \lambda_\alpha)^2}{n}|S|}_{\text{估计误差}} + \underbrace{\|X\beta_S^* - f\|_2^2}_{\text{近似误差}} \right\}.$$

证明

设 $\hat{\beta} := \hat{\beta}_{\text{Lasso}}$，且 $f = X\beta$.在集合 $\|X^T\epsilon\|_\infty \leq \lambda_\alpha$ 上，有：

1. 当 $n|\beta_j| > \lambda + \lambda_\alpha$ 时：

$$n|\hat{\beta}_j - \beta_j| \leq \lambda + \lambda_\alpha.$$

2. 当 $n|\beta_j| \leq \lambda + \lambda_\alpha$ 时：

$$|\hat{\beta}_j - \beta_j| \leq |\beta_j|.$$

因此，至少以概率 $1-\alpha$，满足：

$$\begin{aligned}\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2 &\leq \frac{(\lambda+\lambda_\alpha)^2}{n}\left(\#\{j : n|\beta_j| > \lambda + \lambda_\alpha\}\right) + \sum_{n|\beta_j| \leq \lambda+\lambda_\alpha} n\beta_j^2 \\ &= \min_S \left\{ \frac{(\lambda+\lambda_\alpha)^2}{n}|S| + \|X\beta_S^* - f\|_2^2 \right\}.\end{aligned}$$

对比定理 16.5.1 和普通最小二乘回归

通过与引理 12.3.1 的结果 (iii) 比较，可以看到 Lasso 展示了自动平衡“逼近误差”和“估计误差”的能力，这种特性称为**自适应**（adaptation）.
此外，正则化参数 λ 的选择通常是 $\sqrt{n \log p}$ 量级.因此，为了避免事先知道哪个系数子集是相关的所付出的代价约为 $\log p$，这一代价被认为是较小的.

备注

逼近误差 $\|X\beta_S^* - f\|_2^2$ 实际上包含两部分：

$$\|X\beta_S^* - f\|_2^2 = \|X\beta_S^* - X\beta^*\|_2^2 + \|X\beta^* - f\|_2^2,$$

其中 $X\beta^*$ 是 f 在 X 列空间上的投影.第二项 $\|X\beta^* - f\|_2^2$ 是模型误差，当线性模型完全正确时该误差消失.

推论 16.5.2

假设 $f = X\beta$，且 β 中有 $s := \#\{j : \beta_j \neq 0\}$ 个非零分量.那么，在上述定理条件下，以概率至少 $1-\alpha$，满足：

$$\|X(\hat{\beta}_{\text{Lasso}} - \beta)\|_2^2 \leq \frac{(\lambda + \lambda_\alpha)^2}{n} s.$$

结论解读

上述推论表明，Lasso 估计量对有利情形具有自适应性，例如 β 是稀疏的（即 β 中多数分量为零）.

参数 λ_α 的界定

为了完整分析，需要给出 λ_α 的界限.对于许多误差分布类型，可以取 λ_α 为 $\sqrt{n \log p}$ 的量级.以下以独立同分布的 $\mathcal{N}(0, \sigma^2)$ 噪声为例进行说明.

引理 16.5.3

假设 $Z \sim \mathcal{N}(0,1)$，则对任意 $t > 0$，

$$\mathbb{P}(Z \geq \sqrt{2t}) \leq \exp[-t].$$

证明

对于任意 $u > 0$，有：

$$E \exp[uZ] = \exp\left[\frac{u^2}{2}\right].$$

由切比雪夫不等式得：

$$\mathbb{P}(Z > \sqrt{2t}) \leq \exp\left[\frac{u^2}{2} - u\sqrt{2t}\right].$$

取 $u = \sqrt{2t}$，可得所需结果.

推论 16.5.3

设 $Z_1, \ldots, Z_p$ 为 p 个标准正态随机变量（可能相关）.则对于任意 t，满足：

$$\mathbb{P}\left(\max_{1\le j\le p}|Z_j|\ge\sqrt{2(\log(2p)+t)}\right)\le\exp[-t].$$

证明

由联合概率的上界：

$$\mathbb{P}\left(\bigcup_{j=1}^{p}\{|Z_j|\ge\sqrt{2(\log(2p)+t)}\}\right)\le\sum_{j=1}^{p}\mathbb{P}\left(|Z_j|\ge\sqrt{2(\log(2p)+t)}\right).$$

结合引理 16.5.3 的结果，即得所需结论.

推论 16.5.4

设 $\epsilon_1,\ldots,\epsilon_n$ 独立同分布 $\mathcal{N}(0,\sigma^2)$，且 $X=(X_1,\ldots,X_p)$，其中 $X_1,\ldots,X_p$ 为 $\mathbb{R}^n$ 的固定向量，满足 $\|X_j\|_2=n$ 对所有 j.则对于 $0<\alpha<1$，有：

$$\mathbb{P}\left(\|X^T\epsilon\|_\infty\ge\sigma\sqrt{2n(\log(2p/\alpha))}\right)\le\alpha.$$

备注

当 $\alpha=\frac{1}{2}$ 时，以上界给出了 $\|X\hat{\beta}_{\text{Lasso}}-f\|_2^2$ 中值的界限.在高斯误差情形下，可通过“测度集中”理论推导出 $\|X\hat{\beta}_{\text{Lasso}}-f\|_2^2$ 会集中在其中值附近.

第16.6节结论

本章中，我们看到经典统计学的一些概念在现代统计学中依然扮演重要角色，即使参数可能是高维的.经典的最小二乘法依旧占据核心地位，但现在它结合了正则化惩罚项.更广泛地，M 估计（例如最大似然估计）也可以在高维环境中应用，同时引入某种正则化技术.偏差-方差分解依然起着重要作用，例如指导正则化参数的选择.

收缩估计量在高维统计中起着重要作用.这也与第11.4节的结果有关，我们看到在维数大于2的情况下，样本均值是不可改进的，可以通过收缩估计量改进.

复杂性正则化通常可以被视为一种贝叶斯最大后验（MAP）方法.此外，也可以使用后验均值作为估计量等.如今，贝叶斯方法在高维和非参数问题中非常重要且成功.

复杂性正则化通常用于构造**自适应估计量**.自适应估计量模仿了事先知道目标复杂度的情景.为了评估自适应估计量的性能，通常以目标复杂度已知的情况作为基准.这个基准通常来源于经典统计理论.

第17章 文献

以下是一些与本课程相关的重要参考文献：

- J.O. Berger (1985)
  《统计决策理论与贝叶斯分析》
  Springer出版.这是一本关于贝叶斯理论的基础书籍.
- P.J. Bickel, K.A. Doksum (2001)
  《数理统计学：基本思想与选题》第一卷（第2版），Prentice Hall出版.
  这本书内容广泛且数学上严谨.
- D.R. Cox and D.V. Hinkley (1974)
  《理论统计学》，Chapman and Hall出版.
  包含了对各种概念及其实用意义的深入讨论，数学推导较为简略.
- A. DasGupta (2011)
  《概率论在统计与机器学习中的应用》，Springer出版.
  包含了所需的概率理论背景.（作者即将出版新书《统计理论：全面课程》）
- J.G. Kalbfleisch (1985)
  《概率与统计推断》第二卷，Springer出版.
  涉及似然方法.
- L.M. Le Cam (1986)
  《统计决策理论中的渐近方法》，Springer出版.
  以非常抽象的方式讨论了决策理论.
- E.L. Lehmann (1983)
  《点估计理论》，Wiley出版.
  一本经典书籍，本课程的部分内容参考了此书.
- E.L. Lehmann (1986)
  《统计假设检验理论》（第2版），Wiley出版.
  是上一部书的配套书籍.
- J.A. Rice (1994)
  《数理统计与数据分析》（第2版），Duxbury Press出版.
  一本更为基础的书籍.
- M.J. Schervish (1995)
  《统计理论》，Springer出版.
  数学上严谨且内容广泛，同时也是一本很好的参考书.
- R.J. Serfling (1980)
  《数理统计的逼近定理》，Wiley出版.
  涉及渐近理论.
- A.W. van der Vaart (1998)
  《渐近统计学》，剑桥大学出版社.
  涉及现代渐近理论及例如半参数理论等内容.
- L. Wasserman (2004)
  《统计学全貌：统计推断简明课程》，Springer出版.
  涵盖数理统计与机器学习中的广泛主题.

备注

1. 在涉及统计模型时，为了区分真参数 θ 和索引参数 ϑ，应该更严格地写作 $P_\theta \in \{P_\vartheta : \vartheta \in \Theta\}$.这种符号区分在理论发展中是必要的.
2. “合理性”的定义需要谨慎.例如，在某些模型验证中，将样本分为训练集和测试集会对样本有不同对待方式.
3. 当 X 为二维时，总变差正则化的定义和应用可能更加复杂，需要结合图像处理等具体应用场景加以研究.

---

1. [1] Here is an example, with $N = 3$:

$$
\begin{aligned}
(z_1, z_2, z_3) &= (5, 6, 4) \\
(r_1, r_2, r_3) &= (2, 3, 1) \\
(q_1, q_2, q_3) &= (3, 1, 2)
\end{aligned}
$$

↩

2. [1] Note that the quantity $L(\theta, T(X))$ is random. Note also that in the notation of risk $R(\theta, T)$, the symbol T stands for the map T. ↩
3. [1] If $|\partial \rho_c / \partial c| \leq H(\cdot)$ where $\mathbb{E}_\theta H(X) < \infty$, then it follows from the dominated convergence theorem that $\partial \left[\mathbb{E}_\theta \rho_c(X)\right] / \partial c = \mathbb{E}_\theta \left[\partial \rho_c(X) / \partial c\right]$ or otherwise put, $\dot{\mathcal{R}}(c) = \mathbb{E}_\theta \psi(X)$. ↩